GLOBAL EUROPE
Basel Papers on Europe in a Global Perspective

No. 123

Teresa Pullano and Peter Cornwell
The Divisive Power of Citizenship

Teresa Pullano
Citizenship, Time and Crisis

Abstract: Citizenship is a specific form of social and political closure, thus defining, in a dialectic relation, the changing "shape" of both political subjects and forms of government/statehood. Looking at citizenship as an institution of political subjectivity and statehood, I would like to raise two questions concerning its relation to (historical) time. How does the dual function of citizenship as a form of social closure and a tool for self-emancipation change in times of crisis? What is the relationship between citizenship and moments of historical ruptures, be it wars, revolutions, or times of political transformation? Does citizenship contribute to shaping the temporal dimension of statehood on the one hand and of moments of collective or individual self-emancipation? Citizenship is simultaneously a discursive and a material structure. The hypothesis I want to bring forward is that it shapes our understanding of historical time according to these dimensions.

Key Words: Citizenship, Time, Temporality, Revolution, Crisis

Teresa Pullano is a researcher at the CNR-IRPPS in Rome, Italy. Her research focuses on policies concerning European citizenship in connection to freedom of movement within the EU. She is interested in analyzing the restructuring of contemporary citizenship in Europe as a specific form and structure of statehood. On these issues, she published the book "La citoyenneté européenne: un espace quasi-étatique" (Paris, Presses de Sciences Po, 2014).

When asked to present a paper for the conference "War, Trade and the Divisive Power of Citizenship" (Europainstitut, Basel, September 22 – 23, 2022), which was the closing event of the SNF project on "The Divisive Power of Citizenship", I immediately thought about writing something about citizenship and historical time. I had the unique opportunity to deal with this topic in front of an interdisciplinary audience, many of whom were historians. I always felt that the reflection on the relationship between citizenship and historical time was not present in the debates in political theory and political philosophy concerning citizenship. I was not interested in understanding how the category of citizenship was shaped under specific historical conditions, or even how it evolved – a subject on which a vast body of literature exists. I was looking for a discussion of the mutual dialectic between the institutions that shape political subjectivity, or citizenship, and a different understanding of what historical time is. Does the category of modern citizenship provide stability to the subjects of democratic societies? Does it wither away in times of societal and political disruption and upheaval? What is left of the relation between citizenship and revolutionary ruptures? Very little is available on these topics in the academic literature, maybe due to their highly interdisciplinary nature.

To approach the question, in the paper I start drafting a definition of citizenship as a category of analysis. I then move toward discussing how specific understandings of citizenship as political subjectivity impose themselves in a given epoch. I then look at the level of the subjects: do they feel at pace with these forms of citizenship as political institutions or do they have a different way of reflecting upon their own individual understanding of political time and how it fits, or does not fit, into the general understanding of the epoch? Is the discrepancy between the individual and collective desires to reshape historical time and the institution of citizenship the engine of ruptures, revolts and finally change in the same category of citizenship? Finally, I reflect upon the link between times of crisis and how they affect the institution of citizenship.

I start with a broad definition of citizenship as the category that defines membership in a political community. As such, citizenship implies the constitution of a political body in which its members share common participation. This community can take different forms, such as the one of the city (urban citizenship), the people of a broader collective (the nation, the state or a supranational political entity), or even humanity (cosmopolitan citizenship). In its modern form, since the democratic revolutions of the eighteenth century and especially the "Declaration of the Rights of Man and the Citizen" of 1789 written under the French Revolution, citizenship is tied to the formation of "the people" as the sovereign political subject, in opposition to the power of God or the King under monarchical rule. Modern citizenship thus runs in parallel to the modern

process of secularization of political life, under which political power is no longer an affair of "religious" or "divine" character but is immanent to the body politic of the people. Thus, modern citizenship is associated with two different processes that both have the constitution of the people as the sovereign political subject at its center: the process of democratization, on the one hand, and the constitution of the nation-state, on the other. Citizenship is an institution with a dual character. It is an instrument for struggles towards people's democratic self-ruling and the drive towards freedom and equality (or "equaliberty", according to Étienne Balibar) within the members of a political community. At the same time, citizenship, under modern conditions often in association with nationality, regulates membership within a (national) state and regulates the rights and duties deriving from it as well as demarcating the boundary between those who belong to a (national) state and aliens. The institution of and the struggles for citizenship have an insurrectional dimension and express the striving towards self-emancipation of the individual and collective political subjects. Citizenship, understood in its emancipatory and democratic dimension, is the tool for and the crystallization of civil society's counter-power to the one of government or the state. As such, citizenship indicates the power of the (individual) political subject to shape her public and political life. This emancipatory and empowering dimension of the modern institution of citizenship is always accompanied by the other dimension of citizenship as a tool for governmental power to shape the collective and individual forms of public and private life of the people. Indeed, citizenship becomes a tool to separate the members of a given state from aliens, as well as a means to regulate the participation of the subjects to the life of the state. Citizenship is thus a tool of government, dividing up and differentiating social and political subjects and their rights and duties in a given state. Moreover, I argue that citizenship is an instrument through which statehood, and thus the contours of the political community, are constantly drawn and redrawn. The legal dimension of citizenship and its applications are a powerful tool to shift the spatial and temporal frontiers of statehood.

Starting from this dual dimension of citizenship, I would like to situate it as a pivotal institution in regulating the civil society-state nexus. To do so, we need to develop a dynamic approach to citizenship rather than a static one. Rather than seeing citizenship mainly as a status, defined by certain rights and duties at a given moment and in a given place, I suggest we conceive it as a dynamic form and a political and social process. Citizenship is not a static set of legal provisions, one of the perspectives often adopted by the mainstream literature on the topic.[1] As such, it should be instead analyzed as a specific form of social and political closure, thus defining, in a

---

1   For a critical overview of the mainstream literature on citizenship, see Dimitry Kochenov, *Citizenship* (Cambridge, MA: MIT Press, 2019).

dialectic relation, the changing "shape" of both political subjects and forms of government/statehood. Citizenship is an institution that allows for the challenging of a given form of statehood and political organization in its insurrectional capacity and is a crucial element in the definition of statehood itself. This second aspect is well discussed by Rogers Brubaker, who writes: "(…) the modern state is not simply a territorial organization but a membership organization, an association of citizens".[2]

Suppose we adopt this dynamic approach to citizenship as a social and political process. In that case, the question shifts from "what is citizenship, that is, which rights and duties are attached to it?" to "how does the institution of citizenship contribute to the process of constituting oneself or a group of people as a political subject?" (insurrectional dimension; citizenship as a process of social and political emancipation) and to "how does citizenship act as an essential institution in shaping forms of (self)-government?" (the constitutional dimension of citizenship). Citizenship is thus a form of social and political closure that defines the structure of statehood and acts as a tool for self-emancipation from relations of domination. Looking at citizenship as an institution of political subjectivity and statehood, I would like to raise two questions concerning its relation to (historical) time.

1. How does the dual function of citizenship as a form of social closure and a tool for self-emancipation change in times of crisis? What is the relationship between citizenship and moments of historical ruptures, be it wars, revolutions, or times of political transformation?

2. Does citizenship contribute to shaping the temporal dimension of statehood on the one hand and of moments of collective or individual self-emancipation?

To sketch a path toward answering these questions, I first need to better define citizenship, in general, in historical time. Secondly, I must also tackle the relationship between citizenship and European modernization. Citizenship is simultaneously a discursive and a material structure. The hypothesis I want to bring forward is that it shapes our understanding of historical time according to these dimensions.

---

2   Rogers Brubaker, *Citizenship and Nationhood in France and Germany* (Cambridge, MA: Harvard University Press, 1998), 21.

The Discourse of Citizenship and Its Relation to Historical Time

Citizenship as a Reflexive Institution

As a discursive category, citizenship can be understood as a reflexive institution: it works as a specific grammar through which individual and collective subjects institute themselves as political subjects. There are different ways of thinking about citizenship as a reflexive institution in historical times. Wittrock, following the Basel philosopher Karl Jaspers, argues that "our understanding of history is related to the emergence and institutionalization of forms of critical reflexivity."[3] These forms of critical reflexivity express "the capacity of human beings to reflect upon and to give expression to an image of the world as having the potential of being different from what it was perceived to be here and now."[4] According to Jaspers, the emergence of these institutionalized forms of critical reflexivity inaugurated history as the epoch of reflexive, historical consciousness, or what he names the Axial Age.[5] Could citizenship be understood as such an institutionalized form of critical reflexivity? And if yes, which kinds of historical reflexivity does it bring with it?

More precisely, citizenship has a transformative dimension, which corresponds to what Jaspers named reflexive consciousness: the ability to transcend the immediately given through reason. While the definition of "reason" here can be quite problematic, I would retain the hypothesis that citizenship, as an institutionalized (or not) form of critical reflexivity, has a transformative power that also entails the temporal and historical dimension.

For Jaspers, historical reflexivity also entails a sense of contingency, that is, the awareness of the bounded temporal nature of human experience. To some extent, we could say that this speaks of the relationship between citizenship and secularization. Still, I am interested in a second dimension of thinking of citizenship as a form of the temporal border: to what extent do transformations of citizenship as a form of critical reflexivity indicate transformations in understanding the historical dimension we are in?

---

3   Björn Wittrock, "Social Theory and Global History: The Three Cultural Crystallizations," *Thesis Eleven* 65, no. 1 (2001): 33.

4   Wittrock, "Social Theory and Global History," 33.

5   On this point, see Karl Jaspers, "The Axial Age of Human History: A Base for the Unity of Mankind", *Commentary*, November 1948, accessed on 30 November 2022, https://www.commentary.org/articles/karl-jaspers/the-axial-age-of-human-historya-base-for-the-unity-of-mankind/. In particular, Jaspers writes: "If there does exist such a thing as an axis, or turning point, in history, it must be based on observable or recorded fact; and it must be valid for all men, including Christians. Such an axis would be that point in history where man first discovered the notion of himself that he has realized since, the point in time where there occurred that shaping of man's being which has produced the most important results. And the existence of this turning point would have to be, if not absolutely demonstrable, at least convincing on an empirical basis for Europeans, for Asiatics, and for all men, without the need to appeal to the criterion of a definite religious doctrine. Only thus could it provide a common frame of historical self-understanding.
Such a historical axis, or turning point, seems to be situated in the years around 500 BCE, in the intellectual development that took place between 800 and 200 BCE. There lies, it appears to me, the most crucial turning point in history; it was then that man as he is today was born. Let us, for the sake of brevity, refer to this period as the 'axial age'."

According to Jaspers and then Eisenstadt's theory of the Axial Age, historical reflexivity is associated with macro-institutional transformations within culturally entrenched structural principles.[6] This is, of course, debatable. Still, one of the questions, which is also posed by global history on the one hand and by globalization studies on the other, is the transnational dimension of such forms of critical reflexivity as citizenship and of its transformative power (or not) beyond its specific discursive and material formations.

Still, one key question is: what is the global dimension of citizenship as a form of critical reflexivity? To what extent do changes in the relationship between citizenship and historical time within Europe on the subjective and the governmental levels relate to a global dimension and vice versa?

Citizenship as Simultaneity and Anachronism: The Path Towards Emancipation

Looking at citizenship as an institution of both political subjectivity and forms of statehood, I would like to raise some questions about its historical dimension. First of all, which understanding of time or temporality can we refer to concerning the double dimension of citizenship, that is, its insurrectional or subjective dimension and its governmental or structural dimension? Reflecting upon how transformations of citizenship can mark an understanding of a historical or a larger period, an epoch as in the Axial Age discourse can overshadow contextual and individual differences in the citizenship-time nexus. To what extent is citizenship a tool for subverting the course of time, be it of an epoch (for example, through revolutions or the conquest of larger sets of rights for a larger amount of people) or an individual life?

The French philosopher Jacques Rancière proposes an investigation of the category of "anachronism". First of all, Rancière defines an anachronism as the fact that something happens that would not have been thought possible either at the time or retrospectively. He discusses Lucien Febvre's refutation of Rabelais' atheism in *The Problem of Unbelief in the Sixteenth Century*.[7] One first question is: what is the link between citizenship and unexpected turns of events at a given time and at a given place? This can be thought of either progressively or in a regressive manner. What happens when crises bring about an unexpected and eventually "anachronistic" regression regarding citizenship rights and capacity? What is deemed "possible" at a given time

---

6   Shmuel N. Eisenstadt, "The axial age: The emergence of transcendental visions and the rise of clerics," *European Journal of Sociology* 23, no. 2 (1982): 294–314.
7   Lucien Febvre, The Problem of Unbelief in the Sixteenth Century: The Religion of Rabelais (Cambridge, MA: Harvard University Press, 1982).

and on which grounds?[8] Rancière then deals with a second question: which kind of temporality does history refer to? He treats this question through an elaboration of the categories of time, narration and politics. Different forms of temporality and of rationality are connected to different understandings of historical knowledge. Rancière, following Aristotle, names them as history and poetry. He writes: "The question of the scientific rationality of history is, from the outset, part of the wider problem of poetic rationality, or simply put, fiction. The question of historical knowledge was raised and must be settled on the field of poetics".[9]

Following Aristotle, Rancière argues that poetry is more philosophical than history, which for him meant the mere chronicling of facts. Indeed, "Poetry's superiority is the superiority of one form of temporality over another. History tells the things as they happen in their particularity, one after another, whereas poetry constructs a causal plot that tells us *how* they could happen, and how they happen as a consequence of their very possibility".[10] The kind of temporality we are dealing with here is the second one (poetry), dealing with time as the possible, both in empirical terms and as the result of a specific discourse concerning time and temporality. Citizenship, as one of the conditions of possibility of both political subjectivity and forms of government, is related to historical time as the analysis of the possible. Rancière goes on to argue that "this possibility itself, according to Aristotle, takes on two forms: necessity or verisimilitude."[11]

Both these forms of historical rationality are opposed to what Rancière names the "bad" form of time, that is "the time of mere succession in which things happen one after another, as particular and contingent facts – we could say the time of 'real' (not invented) facts".[12] Historical rationality does not depend on verification, but rather on "the specific form of temporality, the poetic form of temporality that links moments in time as causes and effects".[13]

Citizenship, I argue, intervenes precisely in relation to this form of temporality as poetry. Citizenship, in this meaning, is a "form of action and of life" and as such it relates to temporality. Following Rancière: "But the hierarchy opposing two forms of temporal connection on the

---

8   Holocaust denialists are understood by Rancière as the product of an understanding of history limited to what is deemed possible, as he writes: "[...] this radicalization itself (of Holocaust negativists) was the flipside of the common wisdom typical for post-democratic realism which pretended to stick in every way to the 'possible.'" Jacques Rancière, "Anachronism and the Conflict of Times," *Diacritics* 48, no. 2 (2020): 111.

9   Ibid., 112.

10   Rancière, "Anachronism and the Conflict of Times," 112.

11   Ibid., 112.

12   Ibid., 112.

13   Ibid., 112.

horizontal axis of time succession is itself dependent on another hierarchy, which is a hierarchy of forms of activity and ultimately a hierarchy of forms of life".[14] Citizenship is essentially related to temporality. Citizens are those individuals leading a public life, engaged in transforming themselves and the conditions of possibility of their lives, and thus engaged in transforming time itself: "the time of those who imagine ends and connect the means necessary for reaching them, even at the risk of seeing the chain of causes they have initiated bring about effects contrary to what they intended".[15] Passive citizens are those living in the private realm, a realm in which there is no time, meaning there is no attempt to change the course of events: "On the other side there was the time of those who were called passive men, not because they did nothing but because their activity took place within the time of events happening one after the other, with no ends other than those of immediate survival, no errors, no shift in fortune, no form of poetic rationality."[16]

There are different ways of being in time, influencing temporality as a succession of events. Citizenship, as a way of being political, comes into relation with different manners of being in time, that is, of building oneself as a subject in time. In opposition to an understanding of time as a succession of events, or even as a cause-effect rationality, there is an understanding of time as coexistence and simultaneity. As Rancière writes, "[t]he moment is no longer a link in a chain connecting past and future, cause and effect, it is defined as a principle of coexistence between a multiplicity of situations and activities which otherwise follow various temporal lines – slow or fast, linear or cyclical, straight or broken."[17]

The time of coexistence destroys both the hierarchy of time and the distinction between the time of active men or citizens and passive ones: "It is a time no longer made out of successive instances but of moments that penetrate each other and expand in various directions".[18] This is what Rancière calls the time of the everyday, an egalitarian time which is also the subversive time of the proletarians that he discusses in his essay *Proletarian Nights*: "The everyday has been materially changed and symbolically subverted".[19] Citizenship can mean both understandings of time, the time of the public life versus the time of the everyday, private life, as in Arendt,[20] or the subversive time of anonymous acts of citizenship which subvert the course of things, as in Rosa

---

14   Ibid., 113.
15   Rancière, "Anachronism and the Conflict of Times," 113.
16   Ibid., 113.
17   Ibid., 114.
18   Ibid., 115.
19   Jacques Rancière, *Proletarian Nights: The Workers' Dream in Nineteenth-Century France* (London: Verso, 2012), as quoted in Rancière, "Anachronism and the Conflict of Times," 115.
20   Hannah Arendt, *The Human Condition* (Chicago: University of Chicago Press, 1958).

Parks' acts or as in James Baldwin's narrative.[21] This time is conceptualized by Rancière as the time of emancipation: "This is what coexistence ultimately means [...] the time of the men and women who set out to experiment with forms of life, thought, emotion, and passion that their condition forbade".[22]

To what extent can the category of citizenship make sense of time as dispersion, contingency, simultaneity and thus emancipation rather than time as an organic unity in which the parts are subjected to the whole, as for example the understanding of an epoch or even national time? Simultaneity means the presence of several times in the same time, "not only the presence of the past in the present but also the encounter and clash between several lines of temporality".[23]

Citizenship, as an institution and as the discursive and material condensation of forms of life, expresses both temporality as anachronism and simultaneity, that is as the possibility of changing the given course of time, changing what is deemed possible in an epoch, and temporality as the crystallization of an epoch as a common and totalizing form of life. As Rancière writes: "The democratic time of coexistence woven by modern fiction is hijacked by another time, the time of the epoch thought of as a whole. The epoch is not simply a division within the historical continuum, it is a time unit based on the model of the living, organic totality that nineteenth-century sociology had contrasted with the revolutionary rupture of the "social bond".[24]

The emancipatory character of citizenship, or its "divisive power" at a given time, in a moment of historical crisis, is thus closely linked to how citizenship itself is understood as relating to historical time.

---

21   James Baldwin, *Collected Essays* (New York: Library of America, 1998).
22   Rancière, "Anachronism and the Conflict of Times," 116.
23   Ibid., 116.
24   Rancière, "Anachronism and the Conflict of Times," 116–17.

Reshaping Citizenship in Times of Crisis

So far, I have sketched a path toward a theory of the correlation between citizenship and time as political and historical categories. At this point, allow me to get back to the first question I raised at the beginning of the essay. I asked how the dual function of citizenship as a form of social closure and a tool of emancipation changes in times of crisis. Looking at citizenship as a form of social closure, I defined it as a reflexive institution. I also asked to what extent transformations of citizenship as a form of critical reflexivity indicate transformations in understanding the historical dimension we are in.

Time, Event and Crisis

Following Jaspers and Eisenstadt, historical reflexivity is associated with macro-institutional transformations within culturally entrenched structural principles. Citizenship, as a reflexive institution or structure, tends to have a certain stability over time. How do changes happen in a given situation or around specific events that would produce a transformation? A crisis is thus an event that shakes the stability of citizenship as a reflexive social structure and produces a change in the relation between the reflexive institution and historical time. According to Sewell:

> moments of accelerated change, I would argue, are initiated and carried forward by historical events. […] events typically do more than carry out a rearrangement of practices made necessary by gradual and cumulative social change. Historical events tend to transform social relations in ways that could not be fully predicted from the gradual changes that may have made them possible. What makes historical events so important to theorize is that they reshape history, imparting an unforeseen direction to social development and altering the nature of the causal nexus in which social interactions take place.[25]

Events can thus be defined as a sequence of occurrences that transform reflexive institutions and structures. Still, ruptures in the order of things happen constantly, but they do not necessarily determine the crisis and transformation of reflexive institutions. As Sewell argues:

> An occurrence only becomes a historical event, in the sense in which I use the term, when it touches off a chain of occurrences that durably transforms previous structures and practices. This happens above all when a rupture in one structural and spatial location also produces reinforcing ruptures in other locations.[26]

To be transformative, ruptures need to disarticulate the previous structural network, making it difficult to repair and making the rearticulation of the same reflexive structure possible. Historical ruptures reshape reflexive institutions, bringing to light new forms of understanding

---

25   William H. Sewell Jr., "Historical Events as Transformations of Structures: Inventing Revolution at the Bastille," *Theory and Society* 25, no. 6 (1996): 843.
26   Ibid., 843.

for assemblages that could not otherwise be explained using old reflexive institutions, and thus reframing the political and philosophical grammar of an epoch. Sewell, discussing the taking of the Bastille in 1789 as the transformative event that was identified with the revolution itself, explains how "the epoch-making cultural change – the invention of a new and enduring political category [the one of the revolution] – could, therefore, only take place in tandem with practical changes in institutional and military power relations".[27] The relationship between the French Revolution and modern citizenship is well-documented.[28] Still, very little research has been done on how citizenship is changing as a result of revolutions, wars, or other crisis events since then. Of course, the debate on citizenship and globalization is very rich, but it does not take the angle of reflecting on ruptures in citizenship as a reflexive institution. The perspective is rather one of the incremental changes to modern democratic citizenship, from the national model to the multi-level one. There is no discussion on shifts in the structure of modern democratic citizenship itself. Francis Fukuyama's diagnosis of the end of history after the victory of the liberal order[29] seems to have been fully applied to citizenship, a category that cannot be radically transformed anymore.[30]

As Wagner-Pacifici argues, historical events and ruptures like the French Revolution have as their effects the forging of new political subjects, thus the transformation of citizenship as a reflexive institution. To do so, historical subjects across multiple domains must undertake enormous efforts, as well as "the deployment of objects and acts under the frames of icons, speech acts, gestures, symbols, and more".[31] Transforming crisis and ruptures into new temporal and political lines, as well as into new forms of subjectivity, is a rare event. We must ask the question: is there still politics without the possibility of changing the course of time and shaping it into a new beginning?[32]

---

27   Ibid., 853.
28   Rogers Brubaker, "The French Revolution and the Invention of Citizenship," *French Politics and Society* 7, no. 3 (1989): 30–49.
29   Reference to the work of Fukuyama.
30   Yascha Mounk, "The End of History Revisited," *Journal of Democracy* 31, no 1 (2020): 22–35.
31   Robin Wagner-Pacifici, "What is an Event and Are We in One?," *Sociologica* 15, no. 1 (2021): 13.
32   Hannah Arendt, *The Human Condition* (Chicago: University of Chicago Press, 1958).

Bibliography

Arendt, Hannah. *The Human Condition*. Chicago: University of Chicago Press, 1958.

Baldwin, James. *Collected Essays*. New York: Library of America, 1998.

Brubaker, Rogers. *Citizenship and Nationhood in France and Germany*. Cambridge, MA: Harvard University Press, 1998.

Brubaker, Rogers. "The French Revolution and the Invention of Citizenship." *French Politics and Society 7*, no. 3 (1989): 30–49.

Eisenstadt, Shmuel N. "The axial age: The emergence of transcendental visions and the rise of clerics." *European Journal of Sociology 23*, no. 2 (1982): 294–314.

Febvre, Lucien. *The Problem of Unbelief in the Sixteenth Century: The Religion of Rabelais*. Cambridge, MA: Harvard University Press, 1982.

Jaspers, Karl. "The Axial Age of Human History: A Base for the Unity of Mankind." *Commentary*, November 1948, accessed on 30 November 2022. https://www.commentary.org/articles/karl-jaspers/the-axial-age-of-human-historya-base-for-the-unity-of-mankind/.

Kochenov, Dimitry. *Citizenship*. Cambridge, MA: MIT Press, 2019.

Mounk, Yascha. "The End of History Revisited." *Journal of Democracy* 31, no. 1 (2020): 22–35.

Rancière, Jacques. "Anachronism and the Conflict of Times." *Diacritics* 48, no. 2 (2020): 110–24.

Rancière, Jacques. *Proletarian Nights: The Workers' Dream in Nineteenth-Century France*. London: Verso, 2012.

Sewell, William. H. Jr. "Historical Events as Transformations of Structures: Inventing Revolution at the Bastille." *Theory and Society* 25, no. 6 (1996): 841–81.

Wagner-Pacifici, Robin. "What is an Event and Are We in One?" *Sociologica* 15, no. 1 (2021): 11–20.

Wittrock, Björn. "Social Theory and Global History: The Three Cultural Crystallizations." *Thesis Eleven* 65, no. 1 (2001): 27–50.

Peter Cornwell
Method and Subject: Advances Using Digital Methods in Global History

Abstract: Research activities in the social sciences and humanities have traditionally conceived digital outputs in terms of databases and websites. Practically, most of these are still implemented using SQL, table-based, data structures and synchronic web technologies. However, the constant evolution of software means that browser functionality and security arrangements are constantly changing. As a result, few digital research outputs, except published literature, remain accessible for more than a few years. Recent progress with research data infrastructures is described, which has the potential to improve the sustainability of research investments. This article presents new standards-based annotation techniques, developed in the biodiversity community, which have been applied to global history research questions. Open repository software platforms supporting this 'scientific treatment' approach can now generate technology-independent data resources – supporting long-term reuse by the global community. Promoting institutional change to adopt these developments is discussed, so that costs of data stewardship can be made forecast-able.

Keys words: Sustainability, Global History, Biodiversity, Research Data, Scientific Literature

Peter Cornwell is a research fellow at the Institute for East Asian Studies at ENS-Lyon, a visiting fellow at the Institute for European Global Studies in Basel and professor at the Institute for Modern and Contemporary Culture at Westminster University, London. He is also European co-chair of the Research Data Alliance Preservation Tools, Technologies and Policies Group, community lead for OCFL and WADM in the InvenioRDM Consortium and an expert advisory board member of BiCIKL EU Research and Innovation action in Biodiversity 101007492.

Introduction

This article describes the application of digital methods developed in the life sciences to research in global history, but it is also about the growing significance of digital research data and its effective preservation as both a challenge and a subject for historians. Accelerated by the COVID crisis, paper-based source material previously locked in difficult-to-access archives and specialist libraries is being digitized and made available as citable and preservable data resources. Researchers can build upon the time-consuming discovery and digitization work of earlier practitioners, rather than having to gain access to archive and library facilities individually. But these new resources have little in common with existing PDF-based services, many of which remain restricted to small user communities and are often poorly structured – providing scant metadata for discovery and only uneven text search. Techniques are presented here for building fine-grain empirical data resources as copyright-free 'scientific treatments' – which create new opportunities in global history in areas where progress had stalled. Such evidence, assembled digitally from often fragmentary records, is a game-changer, creating the foundation for future machine-assisted interpretation and moving decisively beyond the illusions of accessibility of recent decades.

However, the specter of digital methods casts a shadow over the social sciences and humanities (SSH), in particular, because of sustainability difficulties with research outcomes produced over recent decades. Despite the growing prominence of the FAIR Guiding Principles,[1] in many cases only contestable publications remain accessible in the long-term: underlying data resources and, in particular, analysis data become inaccessible over time. Software applications, which were essential for access to and the maintenance of digital outcomes such as databases and websites, have become obsolete through evolving IT dependencies. In some cases, costly forensic work has been undertaken to redeliver, using up-to-date technical platforms, outputs which became vulnerable within a few years of their launch.[2] Those projects' investigators had clearly not taken

---

1   The 'FAIR Guiding Principles for scientific data management and stewardship' were established in 2016 to provide guidelines to improve the Find-ability, Accessibility, Interoperability, and Reuse of digital assets. https://www.go-fair.org/fair-principles/.

2   For example, redelivery of research using annotation of manuscripts at Bodleian Libraries Oxford and at Heidelberg University, which had been made vulnerable through obsolescence of web technologies, was presented at International Data Week (IDW), Gaborone | 2018, 10.5281/zenodo.2633630. In the current article, data resources are cited via Persistent IDentifier (PID). PIDs, or 'Handles', which are increasingly being employed to preserve annotation investments, accommodate internet changes over time which otherwise lead to URLs becoming invalid. The Handle system, first developed in 1994 in the U.S. was incorporated into internet governance in 2014 through creation of the DONA Foundation in Geneva, which oversees Multi-Primary Administrator services (MPAs) in countries including China, Germany, the Russian Federation, in the U.S. and via the Smart Africa Alliance. PIDs can be resolved to obtain current URLs by prepending a Handle service, such as https://doi.org/. (The DOI Foundation is also a DONA MPA). For example, the Divisive Power of Citizenship dataset with PID 20.500.14202/hasdai.bqp4m-ax2hw can be reached via https://doi.org/20.500.14202/hasdai.bqp4m-ax2hw while the IDW reference above can be reached as https://doi.org/10.5281/zenodo.2633630.

effective precautions against future support difficulties. In practice, such retrospective funding applications are invariably declined and research investment simply evaporates. This situation is remarkable because, although data losses in the humanities have been particularly acute, systemic sustainability failures affect most scientific domains to some degree. Obdurate policy and institutional shortcomings – no less than technical hurdles – underlie a silent research data sustainability problem which undermines FAIR initiatives. Research leading to improved outcomes with existential climate, epidemiological and antimicrobial resistance challenges relies increasingly upon digital methodologies. Politicization of antiviral research developments during 2020 suggests that the attention not only of philosophers of science but future global historians, will come to bear on research data management over the next decade.[3]

A radical rethinking of policies and infrastructure for long-term access to research data is essential to preventing new losses of investment if emerging digital techniques are applied in SSH. Notably, progress in the biodiversity community, leading to the scientific treatment methodology presented here, is due in part to intergovernmental investments such as CERN[4] and GBIF[5] and also to independent funders such as Arcadia Fund.[6] In contrast, programs such as DARIAH[7] have not provided effective infrastructure for SSH practitioners, despite their original mandates. Rather, international consortia initially assembled by individual researchers and companies have contributed significantly to progress with research data preservation. By establishing models for self-governance, these consortia have become effective at managing sustainability and standards for their technologies in the long-term. Supported largely in-kind by universities and research organizations, these ad-hoc consortia instead of national or European agencies have driven the creation of infrastructure components now being deployed across multiple scientific domains.

Building complete data preservation infrastructure solutions using consortia-provided technology components has been possible since 2021, but widespread uptake by the research community will require decisive action by national and European agencies. Background to these

---

3   Marcus Popplow, "Technology and technical knowledge in the debate about the 'great divergence'," *Artefact* 4 (2016): 275–85, 10.4000/artefact.485.
4   *CERN*, is the European Organization for Nuclear Research, founded in 1954 and funded by 23 Member States. https://home.cern/.
5   GBIF, is the Global Biodiversity Information Facility, founded in 2001 and funded by 41 Voting States and 21 Associate Country Participants. https://www.gbif.org/.
6   The Arcadia Fund is a UK charity established in 2001 by Lisbet Rausing and Peter Baldwin. https://www.arcadiafund.org.uk/.
7   DARIAH – Digital Research Infrastructure for the Arts and Humanities – is a Central European Research Infrastructure Consortium founded in 2006. https://www.dariah.eu/.

circumstances is provided in this article – supporting the case for use of the new digital methodologies described within SSH, and avoiding repetition of its past data preservation failures. At the same time, increasing scrutiny by historians of the research data crisis is also anticipated.

Research Data Preservation Challenges

Little improvement in preservation of research data across multiple domains can be traced since the assessment of Vines et al., published in *Current Biology* in 2013, which found that "80% of raw research data is lost within two decades."[8] It is reasonable to assume that in the intervening years, the volume of research data produced has grown at least as quickly as rises precipitated by internet commerce, mass digital imagery and social media. Approximately 2ZB of data (one zettabyte is a million petabytes, in decimal) was "created, captured, copied, and consumed" worldwide in 2010,[9] and different estimates of data in circulation in 2022, for example of 94ZB and 97ZB, are relatively consistent. However, it is notable that statistics for volumes of research data being generated, in contrast to digitally-originated scientific literature, are more difficult to find. Fundamental reasons for this include a lack of agreement about what constitutes data that should be preserved, and poor understanding about structuring it to enable effective use in the future by others. Raw numerical datasets generated by instruments and computer recognition applications are frequently discarded after analysis products are generated from them because of on-going storage costs. This precludes the possibility of conducting further analyses. Attempts to classify and record procedures and experimental equipment through formal methods, and to preserve multiple versions of analysis software that were employed, are in their infancy. Interim datasets and process-specific parameters and settings, which frequently evolve over the life of a project (and which are essential for reproducibility), remain on research assistants' portable devices and, together with unpublished documents and notes, rapidly disappear from the radar of project managers when personnel are re-assigned or research teams are disbanded. This evaporation remains largely unreported and, even within communities such as the Research Data Alliance (RDA), the "Analysis Preservation" agenda has limited scope and constituency. Where such interim data ceases to be available to future research communities, whether because of technical obsolescence or failure of research activities to have effectively captured it, then the standing of results is compromised. A likely hundred-fold increase in the volume of research data

---

8   Timothy H. Vines et al. "The Availability of Research Data Declines Rapidly with Article Age," *Current Biology* 24, no. 1 (2014): 94–97.
9   Blend Berisha and Endrit Mëziu, "Big Data Analytics in Cloud Computing: An overview" (seminar paper, University of Pristina, 2021), 10.13140/RG.2.2.26606.95048.

being lost over the last decade is serious, because a significant proportion of research is publicly funded, and persistent loss of outputs on this scale represents a failure of public trust. Preserving research data creates the foundation for future work, and it is crucial for later independent verification of research outputs – otherwise, only contestable publications remain.

Scientific treatment examples presented in this article demonstrate that technical solutions for long-term preservation of research data are already available and that, moreover, it is possible to create a new class of robust data resources on which future practitioners can build. However, availability of standards-based infrastructure permitting this is very recent and the prospects for precipitating wider change in data preservation behavior, even within the field of global history, are uncertain.

Several distinct mechanisms contribute to loss of research data – not least, failures of stewardship due to infrastructure and organizational change occurring after effective preservation has been accomplished. Though a comprehensive examination is not attempted here, one aspect – the importance of gathering and preserving interim datasets and records of research procedures using future-proof representations – is critical for practitioners. Slippage intrinsic to the maintenance of information technologies means that, if data is not captured effectively at the time it is produced, the probability of effective preservation increasingly diminishes. Software applications, which are required to manage different research data types, including documents, imagery and numerical quantities, as well as for linking metadata to the digital assets they describe, do not remain static over extended periods. It may not be possible twelve months after it was generated, to run the same version of an application that was originally employed in order to process the research data in the same way. Software applications depend on an ecosystem of other software, which evolves constantly under external engineering and market pressures, leading to an intricate and dynamic web of dependencies. In order to combat this slippage, research data must be exported using non-proprietary (and, where meaningful, standards-based) representations, which are judged to be dependency-free. While not guaranteeing effective preservation, this is an important first step and it makes data 'agnostic' to ever-changing software applications. But for it to be employed efficiently using future technical platforms without costly forensic engineering, further processing is essential – a second step of packaging data for preservation, which is addressed in the next section.

The need for planning and transformation of data to overcome application software dependency hazards; indeed, the requirement to preserve interim analysis data at all, is not well understood

in the research community. Yet responsibility for preservation falls on researchers, since institutions and funding agencies do not provide effective technical guidelines, infrastructures, policies or tools. Focusing on state-of-the-art software applications for their research topic, most practitioners who have adopted digital methods are left to address complex sustainability challenges alone. Few options are open to researchers other than to replicate trees of heterogeneous files from active research IT infrastructures onto long-term storage. But such bulk data 'dumping' does not guarantee future reuse: even if an effort has been made to construct metadata describing these files, compatibility with future software applications is unlikely. Attempting to prolong the use of existing applications is also problematic: in many cases applications are proprietary or based on open-source software which is maintained by fugitive communities. The latter rely on the expertise and goodwill of individuals unless there is strong community governance and a funding stream available to develop and maintain a cohort of experienced engineers. Continued payment of license fees or actively participating in open-source software developments are then the only recourse – but this is not sustainable for more than a few years. Unless research data has been made agnostic, enabling it to be reused with new applications, it is destined to be lost.

A Brief History of Research Data Repositories

Ensuring that research data outputs can be reused with different, possibly yet-to-be-developed technologies in the future is a key task of research data management. Where this requires conversion of the way data is represented, for example by exporting in a technology-agnostic format from a proprietary software application, then a clear transition in status from active research to preservation is established. However, even if data produced in the course of research is intrinsically preservable, attention to making it traceable is still essential. Guaranteeing fixity, that is, confidence that data files of hundreds of thousands of Mbytes remain available with zero corruption for future users, is to no avail without certainty about their identity. This requires that description of the contents of such files, together with their relationship to the research activity in which they were generated, is captured as metadata. Metadata files have to be linked to their referents so that they cannot be separated from them, regardless of data management decisions in later stewardship. If metadata is not available to future users, even carefully-preserved data files at best require forensic analysis before they can be utilized. As time elapses, it becomes increasingly difficult to test for potential software applications that might have been used to create digital outputs, or their significance to the original research, and their potential value evaporates. It is undesirable to integrate metadata with data files themselves, because of the impact on the scalability of discovery and maintenance. The process of creation and management of metadata

to permit effective future discovery and use of research data is a key function of a class of digital infrastructure components called research data repositories. While data is being generated, transferred and processed, it is referred to as 'in motion'. As soon as research activities deem that data has become static – even if succeeded by later versions – its status 'at rest' should trigger commitment to a data repository for secure long-term management. The emergence of pervasive configuration management repositories such as Git[10] in the software development community and, in particular, increasing use of its GitHub commercial instance across a wide range of scientific domains, has influenced behavior in respect of managing research data at rest. However, neither GitHub nor Google's Workspace encourage research practitioners to persist data managed using their products into independent long-term data preservation infrastructures. Significantly, configuration management repositories promote metadata relating to contributions towards software components, and the management of dependencies between them. They provide little support for provenance metadata relating to external objects and processes that are not software programs. Standards for metadata have emerged through the activities of libraries and government agencies internationally since at least 1965.[11] More recent metadata schemes such as that supported by DataCite[12] are implemented by research data repositories presented below, but detailed discussion of both metadata and persistent identifier (PID) technologies is not addressed in this article.

Repository Software Platforms

Data repositories form a class of digital infrastructure components which support the management of data at rest – providing web-based interfaces for administrators and users, as well as APIs for machine access and external organizations including libraries. Originally developed to disseminate and preserve research *outcomes* – predominately scientific literature in digital form and more recently also key datasets – the data repositories currently in use include proprietary services (which are not addressed here), as well as repositories based on multiple open source software platforms supported by international consortia. These consortia comprise commercial enterprises, research organizations and universities. Repository 'instances', based on the shared software platform, are tailored to the individual requirements of the member organizations

---

10   Git is a distributed version control system (software configuration management – SCM) originally used for coordinating work among programmers. It was developed by Junio Hamano and Linus Torvalds and first released in 2005. https://git-scm.org. GitHub is a commercial instance of Git, with more than 100M users. Its parent company is Microsoft Corporation which purchased GitHub in 2018. https://github.com/.
11   Computer scientist Henriette Avram developed the machine-readable cataloging standard (MARC) from 1965.
12   DataCite was founded by organizations from 6 countries in 2009, and 5 additional organizations were approved by the International Council for Science in 2010. https://datacite.org/.

running them. Repository instances can form 'generalist' repositories – for example, supporting multiple faculties at a university, or 'corpus' repositories, addressing more specific requirements of individual research activities and special collections. Because they can be configured more closely for content with a common focus, corpus repositories are able to provide more support for those data types than generalist repositories. For example, by implementing domain-specific metadata and search models, it is possible to offer discovery based on the contents of records, as well as generic terms such as author and date of creation. Significantly, corpus repositories are also able to allocate more resources individually to fewer records. Underlying IT constraints limit maximum data capacity and the performance available for searching repository records in response to concurrent user queries. In contrast, generalist repositories must potentially support millions of heterogeneous records, and therefore limit the size of data files associated with each record, as well as implementing a data model common to all of them. This means that interim and analysis research data, as well as final outcome publications and specific datasets can be preserved routinely using corpus repositories. Nevertheless, consortium member organizations are able to employ the shared repository software platform for both corpus and generalist repository requirements. Members collaborate on maintenance necessary to address changing development environments and operating system dependencies, upon compliance with standards and the network security landscape, and on developing new functionality. All organizations benefit from shared planning and testing of this engineering work. The alternative – of each organization developing and maintaining different repository software – is untenable.

Consortia have created multiple repository software platforms, which have been in service for several decades, but in the last five years, the number of smaller organizations adopting these software platforms has increased. For example, DSPACE repository technology has its origins in work at MIT and Hewlett-Packard Labs before 2002. In 2009, DSPACE joined with the Fedora Commons[13] organization to form not-for-profit DuraSpace. Fedora repository technology was originally developed by Cornell University's Digital Library Research Group in 1997, which together with The University of Virginia established the Fedora Repository project – later, Fedora Commons. In 2019, DuraSpace merged with LYRASIS, an umbrella group of more than 1000 libraries in 28 countries, first established in 1936. The Samvera Community also uses the Fedora software platform, but it utilizes Blacklight and Solr discovery technologies and the Library of Congress's Metadata Object

---

13   Fedora (Flexible Extensible Digital Object Repository Architecture) is a digital asset management repository architecture for building institutional repositories, digital archives, and digital library systems. https://fedora.lyrasis.org. This is distinct from The Fedora Project, which is sponsored by RedHat, Inc., a software company founded in 1993 and now a subsidiary of IBM. https://docs.fedoraproject.org/.

Description Schema (MODS) standard. In contrast, the origins of the Invenio repository platform can be traced to the SPIRES project of CERN and the National Accelerator Laboratory at Stanford University (SLAC) in the 1960s. SPIRES, which continues to operate as iNSPIRE-HEP[14] is a database of particle physics literature, and was one of the first research data resources to be accessible via the World Wide Web. The CERN Document Server infrastructure, based on this experience, was launched in 2000 and moved to the recently-developed Invenio Digital Library platform in 2006.

It is notable that these activities – arising in both the physical sciences and the social sciences and humanities – have converged into two open source software platforms – Fedora under the LYRASIS umbrella and Invenio through the InvenioRDM[15] Consortium. Although these consortia include the largest organizations in their memberships, collaboration is essential for the development and operation of repositories, because of their complexity and large scale, as well as the long-term financial commitment which they demand. Founding-member organizations generally provide financial support in-kind through assignment of their own software development and administrative personnel, or through financial support of external costs such as contract engineers. In some cases, funding agencies have joined consortium governing bodies as observers – sometimes contributing financially at modest levels. Early participation of large institutions has been a factor in other organizations' decisions to join consortia and adopt these software platforms. For example, California Institute of Technology (Caltech) has operated an Invenio-based repository since 2015[16] and Austrian and German universities and Northwestern University Feinberg School of Medicine have contributed significantly to the more recent InvenioRDM Consortium (discussed further below). Smaller, organizations that have joined recently, while they do not provide engineers or funding, nevertheless support valuable testing and internationalization activities. Although protracted, this mode of developing and maintaining complex software responds to the shared recognition by these organizations of the need to retain capacity to manage, make accessible and preserve research data. Relying in their early stages on individuals to precipitate these initiatives while at the same time performing other job functions – and continuing releases of significant new functionality decades later, indicating that development will continue in coming years – repository consortia are not ideal in their approach. However, this self-organizing process demonstrates widespread rejection within the research community of the inadequate functionality and un-forecastable long-term costs of operating proprietary alternatives, and a consortium model is now repeated in the development of other key components of research data infrastructure.

---

14   INSPIRE defines itself as a "community hub" which helps researchers share and find accurate scholarly information in high energy physics. https://inspirehep.net/.
15   InvenioRDM is a turn-key research data management repository software platform. https://inveniosoftware.org/products/rdm/.
16   Library Technology Guides, "Caltech selects TIND Library Management System," *Library Technology Guides*, July 13, 2015, Press Release, https://librarytechnology.org/pr/20852/caltech-selects-tind-library-management-system.

Repository Interoperability Requirements

Repositories are constructed using mature software development environments, and they were not expected to remain in service for more than a few decades before being replaced with repository platforms developed using newer software technologies. Notwithstanding this, investment in developing records using both the Invenio and Fedora software platforms has not been fully reusable when later updated versions of those platforms were released. Significant additional work was necessary to reimplement all of the records in existing repositories when they moved to the new version of the platform, even though attention had been paid to making the data comprising the individual records technology-agnostic. This was because the search and metadata models, as well as logic for access control and underlying database technologies employed by new versions of the repository software platforms, were not compatible with existing implementations.

To address these losses, further development has been required to allow repository records to be used across different existing and future repository software platforms without repeating the investment of creating them. Both repository records' metadata and attached files, and potentially multiple versions of these resources (where they evolve during research processes or through subsequent enrichment) need to be made portable. Requirements for repository interoperability technologies were the subject of an RDA Working Group established in May 2016.[17] A subsequent workshop organized by the Bodleian Digital Library in September 2017[18], which was attended by multiple Fedora repository consortia, proposed creating The Oxford Common File Layout (OCFL) standard – which is based on Unix concepts dating from the late 1960s standardized by the POSIX[19] community. Since 2017 an international consortium, led by Cornell, Emory, Harvard, Oxford and Stanford Universities, has been established to manage, develop and support OCFL, creating important additional opportunities for preservation of research data. OCFL can potentially be used to gather repository records into single, self-contained digital archive objects. The process of making repository records self-describing – also referred to as dereferencing – eliminates the requirement for external information such as schemata, otherwise connected via external links. Such references would otherwise prevent archive objects being employed effectively if the objects they link to no longer exist at a future date. Gathering the complete records of a dereferenced corpus repository using OCFL permits future-proof digital objects to be generated which do not require specific technologies nor infrastructure to make full use of them. Notably,

---

17   RDA Research Data Repository Interoperability WG: https://www.rd-alliance.org/node/50279/case-statement.
18   https://blogs.bodleian.ox.ac.uk/digital/2017/07/06/fedora-and-hydrasamvera-camp-at-oxford-sept-4-8-2017/ and https://ocfl.io/1.1/spec/.
19   Originating in 1988 and now maintained as IEEE Std 1003.1-2017 by a joint group including ISO and IEC.

both the Fedora and InvenioRDM consortia have already implemented preliminary OCFL sup-port.[20] In combination with OCFL, repository platforms can now be used to both manage research data services for global communities via the contemporary internet, and also to archive research corpora for portability between repository software platforms and, critically, for compatibility with future technical platforms.

Very Long-term Preservation

Dereferenced repository records, packaged as archive objects using OCFL, are suitable for offline storage using media designed for long-term access, compared with rotating disk technologies, which have guaranteed operating lifetimes of just a few years. An industry consortium, formed in 2004, comprising Hewlett-Packard Enterprise Company, International Business Machines Corporation and Quantum Corporation[21] collaborates to support Linear Tape Open (LTO) tech-nology. LTO manages the manufacture and on-going development of cartridge-based magnetic tape storage, which has become an IT industry standard – shipping approximately 9.1 million cartridges in 2021. LTO has both an historical track record and a forward roadmap of approxi-mately 20 years each, and its current generation provides 18TB uncompressed cartridge capacity (a terabyte is one thousand gigabytes, or one thousandth of a petabyte), with a 30-year life[22] at approximately 200 euros per unit in small quantities. This enables corpora to be replicated and stored at multiple locations cost-effectively. Critically, LTO creates an 'air-gap' – protecting re-pository archives from online security hazards – as well as bestowing resilience by simplifying automated management of copies of corpora, in geographical locations with orthogonal disaster threats, through robotic tape libraries.

---

20   CERN and Data Futures GmbH demonstrated export for InvenioRDM corpus repositories during 2021, and developed software library support for the Python community. https://pypi.org/project/ocflcore/.
21   Seagate Technology was an early original LTO consortium partner, but its magnetic tape technology was bought by Quantum Corporation in 2004.
22   Recovery of data reliably from LTO depends on storage temperature and humidity, as well as usage. Cartridges ac-cessed frequently by robotic libraries will have lower lifetimes than those shelved in controlled environments. However, the automated replication of LTO contents onto contemporary generation cartridges on a periodic basis, which is effec-tively lossless when multiple copies are maintained at different locations, means that large repository archive objects can be stored with high reliability and at low cost for long periods.

Improving Research Data Infrastructure for Global Historians

Although the trajectories of some of these consortia originate in the 20th century, it is only in recent years that adopting open source repository software platforms has become practical for most organizations. For institutions, the cost of developing and maintaining groups of experienced developers has been challenging, while research data infrastructure development is too complex and open-ended for individual practitioners restricted by specific research grant funding. While, for example, Caltech and Cambridge University operate InvenioRDM[23] and DSPACE repository instances, respectively, use of these services is restricted to their institutions' users. Notably, Zenodo – the repository of the OpenAIRE program, which is operated by CERN in Geneva[24] – accepts deposits from other institutions' researchers, and provides preservation guarantees. Zenodo enables any researcher with an ORCID[25] account to create research data records. However, like the Caltech and Cambridge University repositories, Zenodo is a generalist repository instance for research outputs, with record size limitations and general-purpose data and search model implementations. These restrictions are impractical for global historians digitizing archive source materials and producing fine-grain empirical data. Although the developments reported above are promising, the data preservation landscape at the time of writing still presents researchers who employ digital methods, across the humanities, with two principal challenges:

1. research data generated using many existing software applications must be transformed into technology-agnostic representations which do not attract long-term license fees, to enable it to be employed in the future without forensic engineering; and comprehensive metadata describing such data must be created to permit its later discovery

2. preservation infrastructures must be employed which enable complete research corpora – including interim analysis data – to be made portable, so that they can be stored cost-effectively at multiple locations and used with future repository platforms

The approach taken in the SNF *Divisive Power of Citizenship* project (grant 100011_184860/1) has been firstly to rethink the use of specific software applications for analysis, and to overcome the need for the transformation of research data to make it future-proof. Instead, data has been produced which is at the outset technology-agnostic and also standards-based, requiring no further preparation for subsequent use and robust preservation. Secondly, the release by the InvenioRDM Consortium of a Long Term Support (LTS) version of the software platform in

---

23   InvenioRDM v6.0, released in August 2021, was the first release suitable for production services, https://inveniosoftware.org/blog/2021-08-05-inveniordm-lts/.

24   Zenodo – https://about.zenodo.org/ – is the global catch-all repository of the European Commission's OpenAIRE program FAIR data program. OpenAIRE is a Non-Profit Partnership, established in 2018 to promote open scholarly communication infrastructure for European research.

25   ORCID provides a persistent digital identifier (an ORCID iD) which distinguishes researchers. https://orcid.org.

2021[26] enabled the development of corpus repositories which are not restricted by generalist repository scale limitations and, significantly, which have automated migration support between versions of the InvenioRDM. This key LTS development means that OCFL export is not required to migrate repository instances to current versions of the software platform. Since InvenioRDM corpus repositories can be dedicated to specific research activities, they can be tailored in respect of graphical user interface, data model and metadata implementations and search facilities, and can also accept large-scale analysis preservation and interim datasets. The export at a later date of complete corpora as OCFL archive objects, suitable for off-line storage and replication using LTO robotic libraries, means that complete long-term preservation of the research activity can now be achieved cost-effectively.

In order to work with large-scale historic sources in print, and at the same time avoid research data sustainability hazards, the Divisive Power of Citizenship project (DPC) adopted the 'taxonomic treatment' approach developed in the biodiversity community. Already in use in Zenodo's Biodiversity Literature Repository (BLR)[27], this approach uses machine annotation of publications in order to generate concise data describing species observed, such as taxon rank and material citations (specimens). In biodiversity literature it is important to record a spectrum of related information including: dates, locations and the identities of collecting scientists and other contributors who positively identify and catalogue material. Together, these facts, organized in a formal structure, comprise taxonomic treatments. To extend this approach for problems in global history, DPC evaluated the use of standards-based annotation techniques for a range of source materials at the center of its own research questions. Standards-based JSON[28] representations for 'scientific treatments' of this literature were developed, to form technology-agnostic data resources designed to be efficient for historians to employ generally, without reference to specific research questions.

---

26   https://inveniosoftware.org/blog/2021-08-05-inveniordm-lts/.

27   The Biodiversity Literature Repository (BLR) is a Zenodo community dedicated to making open access and FAIR (findable, accessible, interoperable and reusable) the biodiversity information in hundreds of millions of pages of scholarly publications and in specialist libraries. https://biolitrepo.org/.

28   JSON (JavaScript Object Notation) is a lightweight data-interchange format designed to be easy for humans to read and write and for machines to parse and generate.

Annotation and Scientific Treatment of Literature

Benefits of creating standards for the annotation of literature across scientific domains have become increasingly clear since 2010, with the establishment of the Open Annotation Community Group.[29] Progress initially depended on individuals from a wide constituency, including in high energy physics, libraries, medicine[30] and schools, who produced a W3C Community Draft in 2013.[31] Practical workflows permitting creation and reuse of annotations by scholars were enabled significantly by the emergence of IIIF[32], which was developed for collaborative work on digitized visual material in the cultural heritage sector. In particular, support of OADM by the IIIF Mirador-2 project from 2015 onwards[33] enabled redelivery of earlier research in the humanities which had become vulnerable when custom annotation tools became obsolescent.[34] A full W3C Recommendation – WADM – appeared 2017[35], with support for the annotation of a wide range of objects, no longer limited to 2D material such as digitized literature.

In parallel with the evolution of WADM, work on identifying taxonomic information in biodiversity literature, notably by the Swiss non-profit Plazi Association[36], led from 2009 onwards to the semi-automated production of taxonomic treatment data[37] from digitized as well as born-digital publications. Significantly, Plazi established that scientific facts extracted from literature are unencumbered by copyright restrictions which may apply to the full text of individual publications. This legal precedent also applies to scientific literature outside biology, where the term 'taxonomic treatment' has specific meaning (employing Linnaean classification to precisely categorize living organisms) and has far-reaching consequences for accessibility to scientific information reported in copyright publications. Plazi's TreatmentBank supplies taxonomic treatment data not only to BLR, but also to GBIF, where it forms an important component of biodiversity information provided to government agencies and the international research community. However, having developed in parallel with OADM, Plazi's GoldenGate software application, which mines digitized

---

29   https://www.w3.org/community/openannotation/.

30   A founding co-chair of the Open Annotation Community was Paolo Ciccarese, of Massachusetts General Hospital and Harvard Medical School. More recently, one of the most active organizations in the InvenioRDM Consortium has been the Galter Library at Feinberg School of Medicine, Northwestern University.

31   Open Annotation Data Model (OADM) http://www.openannotation.org/spec/core/.

32   https://iiif.io/community/consortium/members/ established in June 2015, IIIF lists 64 registered consortium member institutions at the time of writing, and many other organizations internationally now employ IIIF-based applications.

33   https://iiif.io/event/2015/ghent/, https://www.nga.gov/audio-video/video/iiif/iiif-snydman-winget-6.html.

34   See *Preserving Scientific Annotation: RDA Working Group Meeting*, 10.5281/zenodo.2633630.

35   https://www.w3.org/TR/annotation-model/.

36   Plazi GmbH, Bern CH-036.4.053.720-5 is a Swiss-based non-profit association supporting and promoting the development of persistent and openly accessible digital bio-taxonomic literature. It is cofounder of the Zenodo-based Biodiversity Literature Repository. https://plazi.org/.

37   https://www.researchgate.net/publication/24244035_Taxonomic_information_exchange_and_copyright_The_Plazi_approach.

biodiversity literature for TreatmentBank, produces taxonomic treatments serialized in XML rather than in JSON. While XML is technology-agnostic and is very widely adopted, IIIF, OADM and WADM are commonly serialized in JSON. IIIF provides a means of 'targeting' WADM annotations to specific fragments of digital objects by providing a coordinate system. An important step in developing the scientific treatment approach for social sciences and humanities applications was to be able to identify key fragments of historic documents automatically – at scale. From 2015, DataFutures[38] had been developing techniques similar to the approach of Plazi's GoldenGate software application but instead using OADM and IIIF. In place of a single text mining application, DataFutures employed a range of existing tools, from Named Entity and Optical Character Recognition (OCR) and neural network applications, to manual annotation using Mirador-2, in order to produce common annotation-based representations of key facts depicted in printed material. The Divisive Power of Citizenship project was able to build on this foundation and produce data resources for historians with the same functionality as Plazi's TreatmentBank resources for the biodiversity community. Basing scientific treatment data resources on IIIF and WADM made them intrinsically future-proof, without the requirement for conversion to overcome technology dependencies. In turn, this also demonstrated the potential for employing WADM as a vehicle for taxonomic treatments – making them interoperable with IIIF infrastructures.

In March 2021, a collaboration was established between the National Museum of Natural History in Paris (MNHN) – a member of the Consortium of European Taxonomic Facilities (CETAF) which publishes the European Journal of Taxonomy (EJT) –, Plazi and Data Futures, to transform existing taxonomic treatments of all of EJT's articles into WADM representations. These treatments were previously created using Plazi's GoldenGate software, and their coverage is being extended as EJT continues to publish new articles. The project enabled assessment of the scientific treatment approach, based on IIIF and WADM, more broadly in other scientific domains. EJT is a 'diamond' open access journal (though key gains translate equally to literature which is subject to copyright restrictions), and so both the full publications as well as scientific treatments could be made freely available. Moreover, EJT comprises a mixture of articles originally appearing in print, as well as born-digital publications. Data Futures generated an InvenioRDM corpus repository with native IIIF services for the EJT articles and developed software to convert GoldenGate taxonomic treatments already available in Plazi's Treatment Bank infrastructure into WADM annotations. InvenioRDM's Mirador-3 viewer then provided an interactive display of both the EJT literature and the taxonomic information originally identified by GoldenGate converted to WADM annotations, together within the corpus repository interface.

---

38   DataFutures GmbH, Leipzig, HRB 38130 is a non-profit company which works on preservation technologies and infrastructures for research data. It is cofounder of the *hasdai* Partnership with CERN.

Taxonomic name ("taxonomicName") annotation derived from EJT article 10.5852/ejt.2015.123 by Plazi's Treatment Bank data resource, displayed here using Mirador-3 in an InvenioRDM corpus repository.

The EJT WADM annotation project, completed before Summer 2021 and available at https://ejt.biodiversity.hasdai.org/, has produced more than 500,000 preservable annotations and demonstrates the feasibility of automatically transforming existing research investments into standards-based data resources. It also contributed to the definition of a new activity, supported by the Arcadia Fund and commenced in 2022, to develop annotation support for Zenodo. Upgrading Zenodo with InvenioRDM functionality in late 2023 will enable display of IIIF-WADM annotations and, moreover, editing and ongoing enrichment of scientific treatments both by experts – for example specialists in particular biodiversity fields – and also for discovery and consumption by machine applications. Significantly, this also creates new possibilities for discovery based on annotations against scientific literature attached to repository records. Repository platforms currently only support mechanisms for searching keys defined in record metadata. The ability to discover records based on the *contents* of scientific literature would represent a significant advance in functionality.

These developments have already transformed biodiversity research methods. Plazi has generated almost a million literature treatment records in Zenodo, which have been processed by the Global Biodiversity Information Facility (GBIF) and the Swiss Institute of Bioinformatics Literature Services[39] (SIBiLS), leading to new publications across the life sciences. Critical cross-domain research, for example, on understanding relationships between habitat loss and virus mutation, which depends on improving automation for analysis at very large scales, has benefited significantly, as demonstrated during the COVID crisis[40], since access difficulties with historic sources are radically reduced. The biodiversity community has been at the forefront of this development because of the circumstances of species loss: assessment of human impact on populations relies on existing publications. While large numbers of new species are still being reported, it is not possible to assess rates of loss of known species effectively through observation. Information about most populations, published by scientists in the intervening years since Linnaeus System Naturae, is still locked in specialist libraries and behind publishers' paywalls: taxonomic treatments release this data for automated analysis.[41]

In January 2023 CETAF announced adoption of IIIF across the spectrum of its publishing activities.[42] On the one hand, standards for scientific annotation were originally driven by a broad constituency in which the medical and physical sciences were prominent. IIIF applications such as Mirador-3 (which has native WADM support) were developed by libraries and social sciences and humanities communities, and their availability has precipitated the transformation of taxonomic treatments of biodiversity literature to be easily editable and interoperable. On the other hand, development of large-scale data resources, such as TreatmentBank in biodiversity, has provided a model for the creation of scientific treatment data resources for global historians. An oscillating pattern of innovation is evident, in which the humanities and life sciences communities, in particular, have in turn led innovation and also adopted standards emerging from each others' developments. The scientific treatment approach is a conspicuous outcome of this synergy: concise data 'blobs', generated at scale from scientific publications, can now be packaged using the vehicle of standards-based annotation, creating a number of important opportunities:

---

39   SIB Literature Services (SIBiLS) is an independent scientific foundation providing tailored information retrieval in biomedical literature for academics, clinical and industry partners. https://www.expasy.org/.

40   https://zenodo.org/communities/?p=covid.

41   It is currently estimated that 500 million literature pages of biodiversity records must be addressed.

42   The Information Science & Technology Commission of CETAF approved the use of IIIF as a standard way for sharing images of natural history objects on 24th January 2023. https://cetaf.org/elementor-7894/.

- packaged as individual data blobs, scientific treatments can be made computationally self-describable: linked via PIDs to source material in which they originate, and to schemata describing the structure of their contents, but requiring no specific application software in order to employ them;

- JSON provides a particularly efficient way to package such scientific treatments: it is widely accepted, lightweight and compatible with many research infrastructures – both open source and proprietary: a data resource comprising JSON blobs has advantages of portability compared with conventional database technologies;

- concise, document-oriented characteristics make scientific treatment data resources efficient to use at both small and very large scales and, significantly, since they are accessible to both human readers and machine learning systems, they create a robust foundation for future technologies;

- because they eschew prose, otherwise essential in traditional modes of communication in scientific literature, scientific treatments reduce barriers to precise machine translation into other human languages (significant volumes of biodiversity literature are published in Chinese and Spanish, for example);

- derivative analysis products can be generated efficiently from scientific treatment data resources – for example, transcribed records of historic foreign residents in East Asia during the 19th century (see *Asia Directories* example, below) cannot be searched reliably because of editorial inconsistencies, however annotations identifying resident records can be processed automatically into datasets permitting reliable search by surname, occupation etc.;

- technologies such as IIIF and PIDs enable scientific treatments to be linked with the context in the source material where they were discovered – this not only means that the original location of an annotation on the page can be visited interactively for verification and further inspection, but that investment in creating scientific treatments can be preserved reliably.

The preceding discussion has focused on historical literature which has been digitized. Resulting page imagery has been analyzed using software applications to recognize characters (plus, in the case of taxonomic literature, conventions of emphasis – italics – are detected: note the word *Thiratoscirtus* above) and assemble a transcription. In turn, the transcription has been analyzed to identify scientific facts. However, new publications accepted by the European Journal of Taxonomy are originated digitally, and the step of generating text for computer analysis from page imagery is unnecessary. Nevertheless, the scientific treatment approach still requires that the context – the position – of key information is preserved. Annotation techniques, outlined above for the case of printed material, enable the WADM standard to be employed to make

extracted scientific facts preservable and link them to the original literature. But this also promotes convergence with established technologies, including nano-publications and the Text Encoding Initiative[43] (TEI) already developed for the analysis of digital texts. Moreover, at the time of writing, the second international meeting of a new initiative to improve the interoperability of a wider class of digitally-created texts has been organized by Oxford and Cambridge Universities.[44] The Interoperable Text Framework (ITF) aims to improve reuse and sustainability not only of born-digital scholarly texts, but also those created through news-gathering and social media. This convergence can be expected to extend what can be achieved using scientific treatment in the future, although that trajectory is not considered further in this article.

The Divisive Power of Citizenship Project

Scientific questions addressed by the Divisive Power of Citizenship project are stated in more detail on its websites, including https://www.divisive-power.org/ and https://asia-directories.org. It is notable, however, that while there are also essays reproduced there, and references and links to journal publications and, in particular, graphical user interfaces with which historians can query data resources produced by the project, their function is solely presentation. Rather, these websites do not maintain or even cache research data as part of their implementations; they employ APIs to access scientific treatment data from DCP repositories. Produced using contemporary internet technologies, it is anticipated that these specific websites will become difficult to maintain within five years because of browser evolution and changes addressing internet security vulnerabilities. Project digital outputs – data resources comprising scientific treatments serialized in JSON and accompanied by JSON-Schemata[45] – form part of InvenioRDM corpus repositories. These repository instances are maintained in the long-term via updates provided for the Invenio software platform by the international InvenioRDM Consortium. InvenioRDM provides website-based interfaces for browsing and searching corpus repository records, as well as maintaining user and administrator community privileges and maintaining research records themselves – separately to presentation websites. In addition, they support access by libraries, using the OAI-PMH protocol[46], and provide APIs to their data resources for use by the research

---

43   The Text Encoding Initiative (TEI) is a text-centric community in digital humanities, operating continuously since the 1980s. It became an approved ISO standard in 2012 (ISO 24612). https://tei-c.org/.

44   The Interoperable Text Framework (ITF) addresses the multiplicity of different formats employed to represent born-digital texts, which make them hard to reuse. 10.17605/OSF.IO/R78GX.

45   https://json-schema.org/.

46   DCP corpus repositories are data resource 'Providers' via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). PMH is an internet protocol that exposes structured metadata for external consuming applications, which is part of the InvenioRDM software platform. https://www.openarchives.org/pmh/.

community and presentation websites, such as https://www.divisive-power.org/. This means that presentation websites always retrieve the most up-to-date scientific treatment data directly from DCP corpus repositories for display in response to internet queries from the research community. Also, replacing presentation websites with implementations using new interaction technologies can be achieved cost-effectively when their support becomes impractical.

An important goal of DPC was the evaluation of the potential for extending the taxonomic treatment approach in use in the biodiversity community to historical sources of value to the project's scientific questions. Analysis of three printed sources in particular informed application of the scientific treatment approach in DPC:

- *Vietnam-Information* newsletters – a series of publications of the Democratic Republic of Vietnam from the 1940s, which were directed at the international community.
- Foreign resident listings and treaties reprinted in *The Asia Directories and Chronicles*, published by the Hong Kong Press between 1864 and 1941.
- American, French and Swiss civil and military archives, which cataloged internees – especially relating to those detained in Indochina during World War II.

These documents comprise a range of printed material, from high-quality impressions on paper in good condition, to those on poorer paper exhibiting 'print-through', and badly-deteriorated low-quality print runs, as well as individual typed sheets corrected and extended with handwriting from archives. Despite presenting different challenges to extracting reliable information, however, these documents all contained information about historic persons and organizations, usually accompanied by locations and dates – amounting to approximately one million 'instances'. There was replication of such individuals – *The Asia Directories*, for example (see below), was published annually and sought in its foreign resident listings to record every European and American present each year in multiple territories. Without reference to external corroborating information, occurrences in these documents alone were generally insufficient as evidence of discovering a unique person or other entity globally. Nevertheless, if this inconsistently-organized printed information could be structured according to uniform historic person and organization schemata as scientific treatments, then data resources could be constructed at scale, which could be queried efficiently for a wide range of purposes. Printed records often included the occupations and residence addresses of persons and employers, although qualifying information such as city and province or business registration usually had to be inferred from document and sub-section metadata. Historic document instance schemata were therefore refined to accommodate this range of information – to avoid having to repeat analyses for multiple purposes – permitting

the creation of general-purpose scientific treatment data resources from the printed materials. Adapting the taxonomic treatment techniques pioneered in biodiversity in this way produced fine-grain empirical evidence from surviving sources describing persons, companies and organizations; potentially overcoming the existing vacuum of traditional archival material about international communities in East Asia during this period.

Three processing stages were applied to all of the document types outlined above: i) identification of annotation target regions, followed by ii) recognition of the characters and words in the text fragments encompassed, thereby forming computer-readable transcriptions, and iii) organization of transcribed text according to common historical schemata employing lightweight JSON Schema standards. The resulting scientific treatment data resources could then be employed efficiently in a range of research activities, not only within current digital infrastructures, but also using future technical platforms.

Creating and Employing Scientific Treatment Data

Structural differences among the DPC source materials meant that multiple strategies had to be adopted to generate scientific treatments from them. Definition of annotation target regions manually by human contributors was necessary where current software applications were unable to identify important document sections reliably. This process, referred to as 'segmentation' was employed to create metadata determining document page ranges for subsequent automated analysis. In contrast, tabulated information and layout in print using regular subsections and columns permitted semi- and fully-automated processing using existing software applications or by developing new reusable software modules. Individual workflows were configured for each source document type to identify annotation targets, recognize computer text from printed characters (where necessary correcting recognition errors or referring to external dictionaries and other authorities to resolve ambiguities) and organize these text fragments into scientific treatment candidates. This process of abstraction bestowed a common structure on information extracted from the different source materials, making instances of persons and companies – occurrences in the historical literature – into interoperable JSON blobs. They could be searched using common methods and consumed, based on their schemata, by external application software such as network analysis tools and formed into scientific treatment datasets for each document. This approach enabled production at the outset of technology-agnostic research data resources, and removed the requirement to maintain software applications employed for analysis, or custom workflow engineering, after completion of DPC's research activities. Data management

planning was made concrete as a result, and financial uncertainties otherwise associated with guaranteeing long-term sustainability were eliminated.

Annotating the News Service of the Democratic Republic of Vietnam

Workflows relating to each of the DPC document types are summarized below, and two sub-projects – listings of the foreign resident populations reported in the *Asia Directories* and archive material relating to internees in Indochina – are described with use case examples. These latter source materials were gathered by the Institute for European Global Studies, Basel (EIB), host of the DPC project. However, documents with which DPC initially tested its scientific treatment were the *Vietnam Information* newsletters, digitized by ENS-Lyon. The objective of this first project was to evaluate collaborative workflows for annotation of historical documents by scholars, and it continued during the spring of 2020, when many researchers were unable to use university facilities because of the COVID pandemic. Accordingly, contributors who developed annotations on the 173 issues of *Vietnam Information* worked remotely because of COVID lockdowns and exclusion from research campuses. Scholars at three institutions participated in this project: The Hesburgh Libraries at Notre Dame University, Indiana, U.S., ENS-Lyon, France and EIB in Switzerland. Digitization of the *Vietnam Information* documents at ENS had been directed towards preservation because of the poor condition of the documents, but neither metadata, nor information about their contents had been produced. The project therefore sought to identify the structure of the publication's issues, and the articles and principal actors addressed therein, by extracting and analyzing masthead information and tables of contents. This was achieved through manual annotation and text entry using Mirador-2. ORCID authentication and contributor workflow management was developed by EIB and task introduction and training for contributors was provided by the Center for Digital Scholarship at Notre Dame. All three institutions provided personnel for the workflows, and the initial annotation task was completed within five weeks, including the creation of an Invenio corpus repository. Providing a dedicated Invenio instance for this project permitted the high-resolution digital imagery produced by ENS-Lyon, as well as annotations and metadata generated from them in this project, to be captured in one corpus repository, overcoming the data size restrictions of 'generalist' repositories such as Zenodo. Since this project preceded the availability of InvenioRDM, which now supports integrated IIIF services and WADM through Mirador-3, the annotations were created using OADM and the earlier

Invenio Framework-3 repository platform[47] was employed. A further work package will migrate the existing Invenio Framework repository to InvenioRDM and convert these annotations and the scientific treatments generated from them to WADM. This will be undertaken during 2023 as part of migration to InvenioRDM, together with other Invenio Framework-3 corpus repositories employing OADM under the *hasdai* program.[48] The *Vietnam Information* repository, which is searchable via the scientific treatment data generated in this project, will continue to be accessible via 20.500.14236/p.v92zj-mep2a.

Foreign Resident Listings Reproduced in *The Asia Directories and Chronicles*

Between 1863 and 1941, The Hong Kong Daily Press published an annual: *The Directory & Chronicle of China, Japan, Korea, Indo-China, Straits Settlements, Malaya, Siam, Netherlands India, Borneo, The Philippines, &c.* (abbreviated hereinafter as *Asia Directories*), which documented the lives of foreigners and activities of foreign companies, associations and their networks with local organizations. It is not known whether all of the physical publications have ever been assembled in one location, but it seems unlikely, because of repeated management changes at the The Hong Kong Daily Press over the extended period of its operations. Furthermore, the complexity and scale even of individual books had been regarded by historians as posing insurmountable challenges to automated analysis as a serial, and its volumes are too daunting for manual progress. The history of this publication, and discussion of its potential for historians as a complete edition is discussed in detail by Herren and Cornwell in Communication at the Dawn of the Golden Age.[49] Gathering information each year for publishing in the Asia Directories was accomplished by individual agents, distributed across vast territories, before radio communication50 or the completion of railway networks in East Asia. As a result, information reported from different geographies bears traces of independent editorial decisions and layout conventions, and the vocabulary adopted in print also evolved as European concessions were established, city names changed and even national boundaries moved.

---

47   Invenio Framework-3 is the version of Invenio, still in service, which succeeded Invenio-v2 after 2015, and upon which Zenodo is currently based. https://invenio.readthedocs.io/en/latest/community/history.html.

48   The *hasdai* Partnership of Data Futures and CERN has operated a network of Invenio-based corpus repositories for biodiversity and SSH research activities since 2018 under a CERN Memorandum of Understanding.

49   Communication at the Dawn of the Golden Age: The Asia Directories and Chronicles. Combining historiography and data-science in a micro-global approach to foreign resident activity in East Asia during the 19th and early 20th centuries. Herren, M. and Cornwell, P. 10.5281/zenodo.5579598.

50   Telegraph services relied on trained operators until after 1914 when automated transmission was developed.

From 2016 onwards, EIB assembled a digital corpus, which at the time of writing includes all the volumes of the *Asia Directorie*s except 1866, 1867, 1872, 1875 and 1884 – currently a total of 110,776 high resolution page images. In contrast to the low-volume printing of *Vietnam Information* publications (above), optical character recognition (OCR) produced computer text from the digital page imagery of the Asia Directories with approximately 96% success for most volumes of the serial that were available. Manual annotation was employed for segmentation of the contiguous pages, to:

- establish correlation between inscribed page numbers (which employed inconsistent alphanumeric conventions in the printed volumes) and monotonically increasing numbers of page image files;
- identify section and sub-section headings and page numbers (since volumes of the serial generally lacked overall tables of contents);
- define page ranges within the volumes which contained specific listings, such as foreign residents, corporations, societies and reprinted treaties.

This assisted with the automation of OCR for specific listings year-upon-year and also, later for navigation and browsing using presentation websites. The tabulated format of foreign resident listings, and headings and sections of company listings and paragraphs of reprinted treaties, enabled the creation of annotation target areas at scale through software analysis of the ALTO-XML[51] output of OCR. Consequently, modest software development by DPC enabled the automated creation of more than 930,000 annotations for foreign residents alone, each comprising coordinates of the respective page image fragments plus the ASCII transcription of the text encapsulated by them. This provided a framework for the production of as many scientific treatment datasets – represented as JSON blobs. WADM annotations, each comprising target definition, plus transcription, plus metadata including a creation timestamp and creator identifier (such as ORCID for manual annotations), plus a research activity identifier – form one part of DPC scientific treatments. Information extracted from the transcribed text corresponding, for example, to historic person instances in the printed source and organized according to respective schemata, was linked to these annotations. Establishing annotation collections enables the interactive display of page image fragments corresponding to individual treatment datasets, using a range of IIIF-compatible viewers such as Mirador-3. Transcribed text strings were analyzed primarily to detect whether they complied immediately with schemata, though in practice exceptional for this source (see *Communication at the Dawn of the Golden Age* for a more detailed description).

---

51   Analyzed Layout and Text Object (ALTO) is an open XML Schema developed by the METAe EU-funded project. ALTO enables description of text and layout information of digitized printed documents.

Instead, lack of adherence of the printed listings to conventions – such as persons' forenames and initials, or honorifics, and irregular descriptions of persons' occupations – required frequent transformation of the transcribed text string. A more significant software development effort was therefore required to build a workflow for 'tokenizing': re-ordering and assigning (and, where necessary, correcting) – the component words of the transcriptions against the specification defined by the schema. Where automated correction was deemed unreliable, candidate treatments were marked for manual resolution by a team of scholars, and completely automated processing was abandoned. Otherwise, tokens of the transcribed text (surname, for example) were structured according to the schema and serialized as independent JSON blobs linked to the WADM annotation from which they were derived. The resulting scientific treatment data resource – continuing here to use the example of the foreign resident listings – for occurrences in the printed source, therefore comprised three components – linked together but consumable by client applications individually:

• WADM annotations serialized as JSON, comprising target definition and transcribed text, plus source document and research activity metadata (via persistent identifiers);

• historic person information, derived from analysis of transcribed text and tokenized against respective schemata, serialized as JSON and linked with source document and research activity metadata (as well as the respective WADM annotation);

• a JSON schema describing a canonical instance of an individual person instance occurring in historical documentation, which insures that the annotation and tokenized datasets are efficient to consume by client applications.

Combining the tokenized text and respective annotation to produce atomic scientific treatments is computationally straightforward, but retaining these three components independently results in greater flexibility. The WADM forms strict annotation collections, with a single 'motivation' – that is, 'transcription', which can be used by a range of IIIF applications. Even if a IIIF service for the Asia Directories does not exist at a future date, it can be recreated automatically from the high-resolution page imagery deposited in the corpus repository. Similarly, tokenized foreign resident, company and other datasets, can be employed, for example, for corroboration with other historical sources to increase the certainty that an instance discovered in the *Asia Directories* relates to a known person entity. Employed together, the components of these scientific treatments enable graphical user interfaces (GUIs) to deliver page image fragments, corresponding, for example, to foreign resident surname instances in a specific year, interactively via IIIF. The same GUIs – and significantly, machine interfaces – can access other historic sources from which scientific treatments have been produced without modification. These possibilities are demonstrated

by the two websites – https://asia-directories.org and https://divisive-power.org – described above. In particular, the latter provides access to multiple archival sources, which differ considerably both in the organization of their materials and in the techniques that were necessary to create scientific treatments. Such cross-corpus interoperability, intrinsic to data resources constructed with scientific treatments, permits 'what-if'-style enquiries by scholars, extending across heterogeneous sources (since diverse scientific treatment schemata can be interpreted dynamically), and new modes of developing and testing hypotheses. This is discussed further in the next use case, which traces internees in Indochina between 1919 and 1945.

Application of scientific treatment of the Asia Directories foreign resident listings can also be demonstrated using geographic visualization. A preparatory dataset produced from the *Asia Directories* in 2018[52] was used to evaluate workflows for tokenization of transcriptions of person instance annotations. Scientific treatments derived from resident listings in the 1896, 1899 and 1934, 1937 volumes were selected as a benchmark for the purpose of developing dictionaries for automating correction of the other volumes of the serial. These years are pivotal in relation to historic events in East Asia. The First Sino-Japanese War, waged from July 1894 to April 1895, was followed by the establishment of large numbers of small communities of foreign residents throughout East Asia. It is recognized that in the 20th century, consolidation of foreign residents in larger communities in coastal cities occurred. This was followed by a marked exodus during escalating conflict in the Second Sino-Japanese War between July 1937 and September 1945, originating in the Japanese invasion of Manchuria in 1931. These population shifts are visible when the benchmark dataset is rendered geographically using the QGIS application[53] (a description PDF, together with datasets and QGIS visualizations for all four years, are available via Zenodo). The 1896 data reproduced here shows large numbers of small groups of foreign residents at locations in the Chinese mainland and around Hanoi, Saigon and the Straits Settlements of Penang, Singapore, Malacca, and Dinding.

In contrast, in 1937 it is evident from the rendering of the treatment datasets that the small foreign resident communities, which were spread across the mainland during earlier years, have been replaced by larger communities along coastlines – even as these populations were diminishing towards the end of the decade before World War II. Establishment of foreign communities in Harbin is an exception to this trend. Analysis of intermediate years – even within the limited Foreign Residents Benchmark Dataset – shows a string of small communities extending from the Dairen

---

52   Asian Directories: Foreign Residents Benchmark Dataset, 2019, 10.5281/zenodo.2580998.
53   QGIS is an international consortium, currently organized as an Association legally based in Switzerland: https://www.qgis.org/.

peninsula (now Dalian) towards Harbin, as a new railway was constructed. These intermediate settlements had vanished by the 1930s, leaving foreign residents in Harbin relatively isolated.
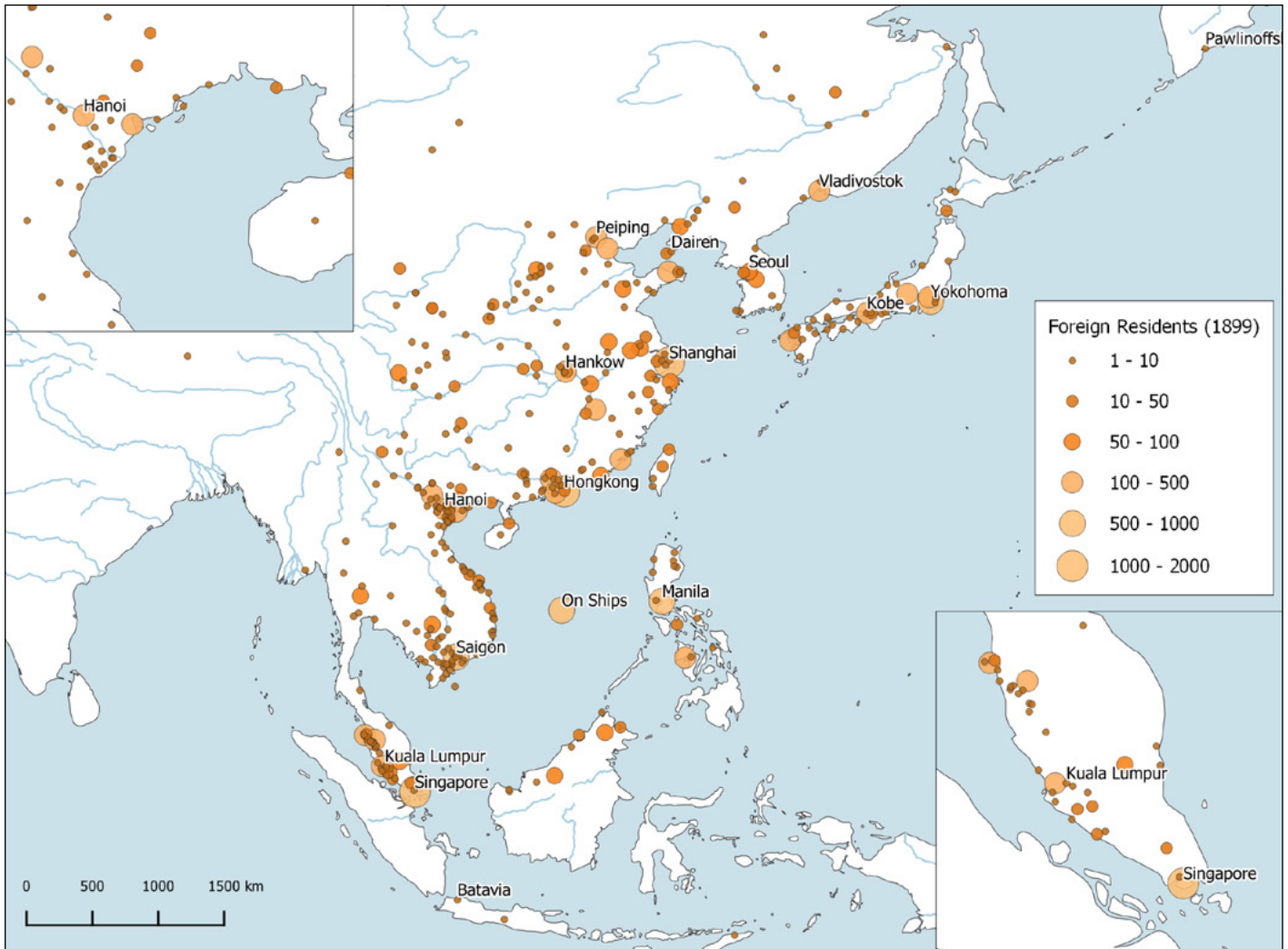
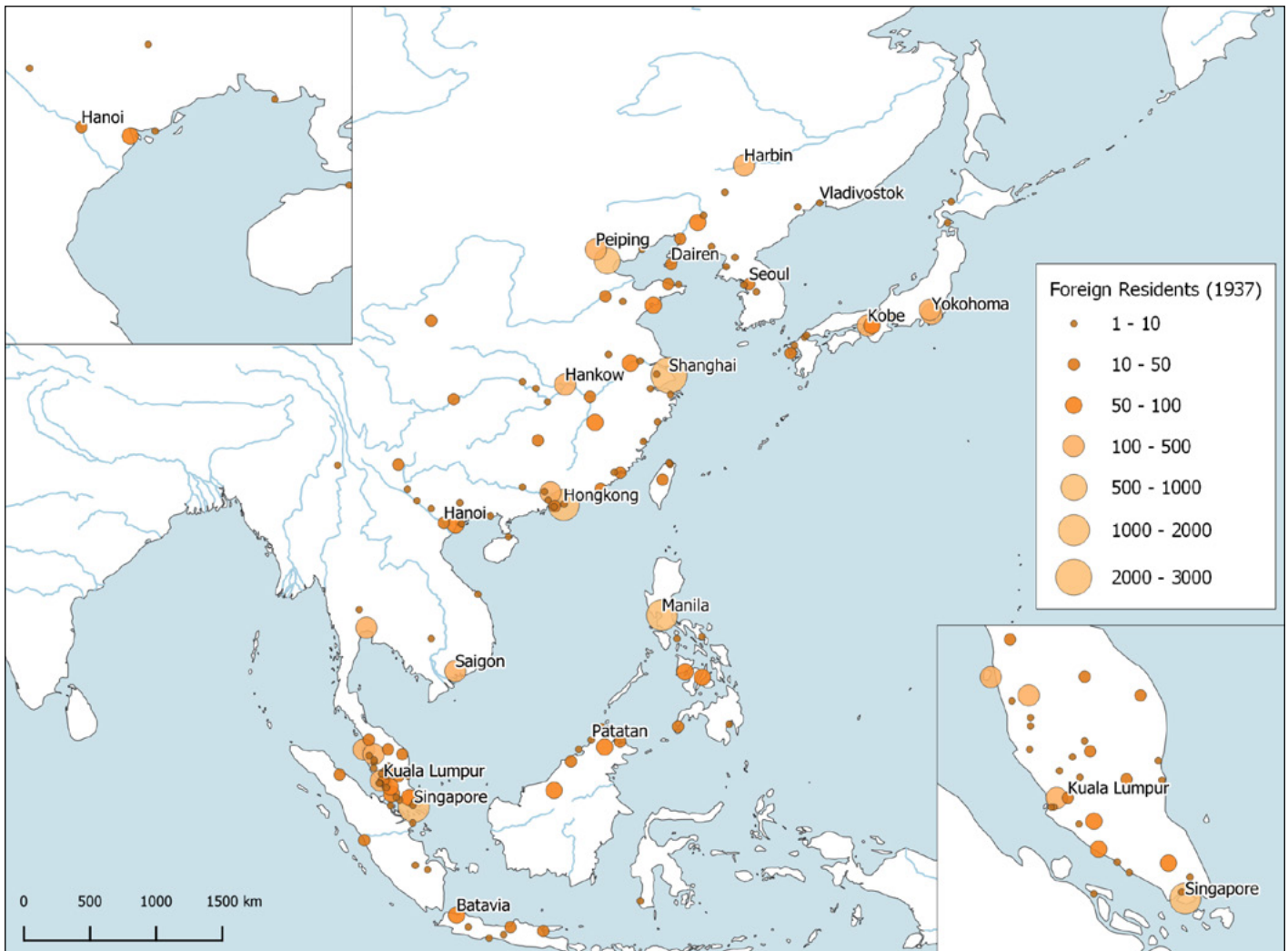Foreign Resident Settlements in East Asia from The Asia Directories and Chronicles, 1899

The QGIS visualizations reveal a different outcome in the Malaysian peninsula where, instead of uniform population movement towards the coasts, the establishment of *new* foreign resident settlements inland is observed, as well as growth of coastal communities. This contrast with the trajectory of foreign communities in other territories was initially unexpected, and does not seem to have been reported elsewhere. However, British interests did not dominate in the Straits, which connect the Indian and the Pacific Oceans. Rather, the vestiges of previous colonial projects confronted one another in this region. Portuguese and Dutch possessions were in close proximity to the newer British, French and U.S. dependencies – and these overlaid more ancient Chinese and Arabic trading routes. Instead of the pervasive influence of the British Empire in the

Straits of Malacca and Singapore throughout the late 19th century, it seems from this empirical data that independent networks of foreign residents may have taken a significant role in shaping this region, and that its development cannot be fully understood as an expression of imperial policies. Rather, a case for the agency of an independent international community might be argued, distinct from the colonial projects of Western states on one hand and later Japanese expansionism on the other hand. Instead of territorial possession, the effectiveness of this community arose from the development over an extended period of local relationships originally established in the 19th century. These small social groups profited from connecting indigenous communities with early channels to major centers of trade. The tiny communities of foreigners discovered on the mainland and depicted in the 1899 visualization might be considered nodes in this *ad hoc* network, and their subsequent relocation to the coast belies a fundamental shift in commercial practice in this region. In contrast, scientific treatment data from the *Asia Directories* supports the view that this international community continued to expand its trade networks in the Malaysian peninsula and Straits Settlements at least until the Second World War.

Enriching Historical Sources

The achievement of the Hong Kong Daily Press in compiling extensive information not only about foreign resident communities year-upon-year, but about many other commercial, national and social enterprises as well, was ground-breaking in many ways. The circumstances of its operation over such an extended period are examined more closely in the *Communication at the Dawn of the Communication Age* article. However, at the time of that publication, the volumes of the *Asia Directories* could not be analyzed as a serial at scale as in the present article. For example, the establishment and growth of communities originating with a few individuals can now be traced over extended periods and their geographic relationships identified at a glance – co-located in many cases with railway developments, road construction and river shipping over thousands of kilometers. In contrast, person instances in the printed material are listed alphabetically, not by location. Newly interpreted using fine-grain person instance data, the *Asia Directories* provide important insight into the importance of water transport before the construction of railways and surfaced roads[54]. Also, in the late 19th century, significant numbers of non-military persons as well as service personnel are listed with addresses on ships (see the 1899 QGIS rendering, above – located, for rendering purposes, near Saigon) while others travelled abroad, although nominally resident in East Asia, more than half a century before widespread access to air travel.

---

54   See Anne Reinhardt, *Navigating Semi-Colonialism. Shipping, Sovereignty, and Nation-Building in China, 1860–1937* (Cambridge, Massachusetts: Harvard University Asia Center, 2018).

Foreign Resident Settlements in East Asia from The Asia Directories and Chronicles,1937

More detail about these individuals' circumstances, compared with the scientific treatment schema employed for these visualizations, is accessible via IIIF, which interactively connects the data relating to each person with the full record on the original page in the printed volume. These advantages of searching, scale of analysis and linking of data resources with the printed source materials, support new insights from material heretofore deemed inaccessible as a serial. In contrast to traditional interpretations and histories of institutions, this micro-global approach succeeds by integrating large numbers of small-scale facts – scientific treatments of historic literature represented as JSON blobs.

Long-term Sustainability

The preliminary data resources presented here were not constructed to answer specific research questions, rather they form a foundation for production of more specific or more sophisticated scientific treatment data in the future. Moreover, returning to the earlier discussion about vulnerability of research investment using digital methods, the approach outlined here has been to produce intrinsically preservable agnostic data at the outset. Visualization using QGIS (which is a Free and Open Source Software – FOSS – application) provides a simple example of analysis using the scientific treatment data from the *Asia Directories*. Making this new data resource accessible via websites tailored for use by the global history community is demonstrated further in the final use case, below. An InvenioRDM corpus repository has been constructed with records comprising digitized pages of all the available volumes of the *Asia Directories*, and including the respective scientific treatment data as downloadable JSON files. This repository provides its own website using technologies maintained by the InvenioRDM consortium, as well as an unrestricted IIIF service for the printed volumes' page imagery, which can be employed by external applications. The latter also permits browsing of the WADM annotations within the repository interface. Depositing all the high resolution page imagery files (many of the print volumes were digitized by the DPC project, and are not available in digital form elsewhere), as well as the scientific treatment data, this corpus repository also serves the long-term digital preservation of this serial. InvenioRDM enables production of an OCFL object file that can be replicated using LTO at multiple locations.

Assembling General-purpose Data Resources

Scarcity and difficulties of access to source materials has been a persistent challenge for global historians working on East Asia. Nevertheless, understanding of the processes at work in the region is of increasing importance because of its complex and rapidly-evolving relations with the West. During the decades after 1900, East Asia was one of the world's most turbulent regions, with colonial conflict, civil wars and revolutionary movements followed by two World Wars. The loss of libraries, archives and other traces of foreign communities continued after World War II through China's Cultural Revolution and The Indochina Wars, waged in Southeast Asia from 1946 to 1991. Today's relationships between the U.S. in particular and Cambodia, China and Korea have global implications. America's war in Vietnam in the 1960s and 70s helped bring to power the Khmer Rouge. Hun Sen, a former Khmer Rouge commander who has been Cambodia's prime minister since 1985, is one of China's closest regional allies. Sen's government and the U.S.

exchanged diplomatic statements in February 2023 over the closure of the Voice of Democracy newspaper and the jailing of the Cambodian opposition leader in advance of national elections.[55]

The DPC project gathered archival materials to support multiple PhD research programs, built collaborations with other global history institutions such as the Institute for East Asian Studies at ENS-Lyon and worked with libraries to digitize chambers of commerce journals, as well as assembling the *Asia Directories* digital corpus. This led to the creation of approximately 1.5 million fine-grain datasets, permitting a practical evaluation of the scientific treatment approach. The heterogeneous nature of these source materials contrasts with the existing use of taxonomic treatment in the biodiversity community. The PhD program of Christian Futter – The Divisive Power of Citizenship and Loyalty (DPCL) sub-project of DPC – employed this methodology to investigate national allegiances and the impact on the citizenship of foreign residents interred in French Indochina from 1929. Electoral rights, residency and tribunal records from Indochina as well as business relationships and commercial proceedings not previously digitized, were gathered at the level of individual persons from multiple small archives internationally. For example, 15,096 persons are documented who lost their citizenship due to legislation introduced during this period, as are the applicants to and proceedings of a commission established to award 'déportés ou d'internés politiques' or 'déportés ou d'internés résistants' status. These materials, organized as digital collections of page images forming records in an InvenioRDM corpus repository, tested the scientific treatment approach further than the *Asia Directories* and *Vietnam Information* sources already discussed, because of their irregular and frequently more complex page layout, varying physical condition and manual amendments.

This example, which shows record 20.500.14202/hasdai.bqp4m-ax2hw of the InvenioRDM corpus repository developed by Futter (https://dpcl.ei-basel.hasdai.org/), accessible also via a Zenodo DOI, was annotated through analysis of ALTO-XML output from OCR and required manual scrutiny. The annotation highlighted in the figure above is accessible to external applications as a page fragment, which can be delivered interactively by the Invenio IIIF service (demonstrated further in the following example). Tokenization of the transcribed text string that forms the body of this WADM annotation produces a scientific treatment comprising surname, forename, and city of residence of each individual together with archival provenance, such as purpose of collation and the date and status of source, for reliable discovery. Street address, occupation and employer are also available in the treatment as unstructured text, for further analysis but not yet reflected in the schema and so not searchable using vocabularies or external authorities (note identification

---

55   Reuters and Agence France-Presse reporting in Phnom Penh, 13th February and 3rd March 2023.

Annotation displayed using Mirador-3 IIIF viewer integrated with InvenioRDM corpus repository, of documents held at the Archives Nationales d'Outre Mer, at the Mairie d'Aix-en-Provence. 10.5281/zenodo.5666615

of 'Saigon' via a city vocabulary in the illustrated example). This treatment data, serialized as JSON, is attached to the digital collection record in the corpus repository as a downloadable file, together with high-resolution page imagery files. The corpus repository contains records corresponding to digitized material from 46 other archival collections of varying scales. These were analyzed to generate scientific treatment datasets using a range of techniques determined by the printed source. Manual annotation and entry by groups of contributors, as described above concerning the Vietnam Information newsletters, was necessary in some cases. For other material, manual correction of annotations generated by software analysis of OCR data was possible, as well as semi-automatic checking of the output of automated analysis. It is notable that custom engineering to accomplish this was restricted to creating workflows to connect existing applications via APIs, and managing permissions for groups of contributors (research assistants and

other investigators) whose intervention was necessary when fully automated processes could not be relied upon. New analysis instruments did not have to be developed, which would otherwise have established a maintenance burden.

Each of the corpus repository records also forms a separate Zenodo deposit (which is also linked from the corpus repository) to maximize accessibility. Such Zenodo records are discoverable via OAI-PMH, and provide respective downloadable scientific treatment files and descriptions of archival sources. In contrast, the corpus repository supports an independent IIIF service as well as downloadable treatment datasets, and it enables preservation of the digitized page imagery. Creation of archive files, by gathering digitized sources document imagery and respective scientific treatment data resources using OCFL, allows interoperability with other repository software platforms, and also near-line replication for long-term storage using LTO libraries.

The goal of producing these data resources was to make available the scientific facts described in historic documentation in a structured form, enabling efficient use in a range of research activities. Crucially, consuming applications do not have to parse raw transcribed text – or detect OCR errors or local editorial conventions and variations. Rather, applications are provided with tokenized occurrences – in this case, instances of individual persons appearing in historic sources – which are defined unambiguously by schemata.

Futter's scientific program, which has produced approximately 50,000 corrected person instance treatments from 46 archival sources, is described in detail in other publications, including by its https://www.divisive-power.org/ presentation website, described above. Additional website and machine-accessible interfaces to DPCL digital outputs are provided by its corpus repository and Zenodo. Both of these infrastructures are maintained in the long term by the InvenioRDM Consortium. The DPCL presentation website provides an interactive search interface with cross-corpus access to multiple data resources generated within the larger DPC project. An example trajectory discovered using such fine grain information about particular individuals is presented below. However, it is notable that the DPCL data resources, excepting only one restricted source, make available all of the investment in locating archival sources, gaining physical access for the organization of specific printed materials and their digitization and then analysis – for other researchers' unrestricted use. Direct access to source materials is often impractical for many researchers, and increasing travel restrictions and political impediments, together with the deteriorating physical condition of physical documents, make the availability of remotely accessible digital equivalents more important still. Moreover, comprehensive capturing by DPCL

of documents as digital collections, together with effective metadata, means that more specific or sophisticated analyses can be applied in the future to the already established data resources.

Tracing Internees in French Indochina between 1929 and 1942

In contrast to the use of bulk scientific treatment data from the *Asia Directories* in the geographic visualization example presented above, it is possible, using treatment datasets developed in DPCL, to trace in more detail the circumstances of specific individuals who became subject to internment policies in French Indochina during the Second World War. A wave of repression directed against citizens of Allied nations in Indochina commenced at the end of 1941. The Japanese had maintained a formidable military presence in the territory following troubled negotiations with French authorities during the Summer of 1940. Following the initial assault by the Wehrmacht and subsequent defeat of France in Europe six weeks later, Japan sought concessions from the French administration. After the attack on Pearl Harbor in December 1941, Japanese pressure on formerly neutral French Indochina increased still further, manifesting in measures against citizens of states then at war with the Empire of Japan. The French authorities in Indochina reluctantly acquiesced, and Allied civilians were interned in the civilian internment camp in Mytho, approximately 70 kilometers outside Saigon, in a decommissioned military barracks. The treatment of inmates, however, was acceptable compared to other Civilian Assembly Centers in Asia – primarily because the French authorities sought to avoid provocation of the Allies, and they presented themselves as a neutral actor. Although reacting to Japanese pressure, this policy of internment was therefore implemented reluctantly by the French administration.[56]

Approximately 188 American, British and Dutch citizens can be traced in Indochina immediately before these measures, but of those, 28 were placed under house arrest and 48 under surveillance; while 50 citizens of Allied states were interned at Mytho.[57] Archival material from the Mytho internment camp concerning the U.S. citizens was compiled by the Swiss Consul in Saigon, Hans Hirsbrunner as "Les ressortissants américains confinés à Mytho" and was sent to the American

---

56   Lenzinger: TELEGRAMME (C), BANGKOK, 20.5.43.8h40., Politique Intérêts, Berne. H. Nr. 39, Bangkok 20.05.1943, BAR, GRANDE BRETAGNE en INDOCHINE, MESURES CONTRE des RESSORTISSANTS, Signature: E2001-02#1000/114#545*.
57   No Author: No Title (commences: Les représentants de l'armée japonaise en Indochine). H. Nr. 36, 06.04.1943, BAR, GRANDE BRETAGNE en INDOCHINE, MESURES CONTRE des RESSORTISSANTS, Signature: E2001-02#1000/114#545*; No Author: No Title (commences: "Il y a actuellement en Indochine"). H. Nr. 21, 06.05.1943, BAR, U.S.A. en INDOCHINE, MESURES CONTRE des RESSORTISSANTS, Signature: E2001-02#1000/113#451*.

delegation in early October 1943. Additionally, a list of Canadian citizens titled "Les ressortissants canadiens confinés à Mytho" was sent to the British delegation. It is likely that both were compiled in preparation for the second American-Japanese exchange of civilian internees. This material resides in the Swiss Federal Archive at Bern and has been digitized by DPCL. Its corpus repository records (20.500.14202/hasdai.lw2cn-53gqz and 20.500.14202/hasdai.7qp5b-nsadg respectively) provide downloadable treatment datasets and IIIF page imagery services. Additionally, 10.5281/zenodo.6813785 and 10.5281/zenodo.7065068 respectively replicate downloadable datasets.

It was productive to attempt to trace persons appearing in Hirsbrunner's list of internees because of an earlier DPC sub-project[58], which had examined American civilian internee records provided by NARA (the U.S. National Archives and Records Administration). Based on Red Cross reports from the region, *Record Group 389* contains information about U.S. military officers and soldiers and U.S. and some Allied civilians who were prisoners of war or internees. The earlier DCP project had noted from the NARA Custodial History Note that the U.S. War Department used punched cards to manage this information, although "[t]he punch card records were transferred to NARA with virtually no agency documentation" and had subsequently been digitized. Probably related to these events, this NARA Record Group[59] displays likely corruption due to the various transformations it has undergone. The DPC project developed a scientific treatment of NARA 389, producing a JSON dataset and schema, which rectified these difficulties for 126,207 of the 143,374 person records in the original Record Group. Consequently, it is possible to discover in the new treatment of NARA 389 a Swiss-born American, Gabriel Denis Corvissiano, who was interned in Mytho. DPCL's Mytho dataset confirms that Corvissiano was present, together with his wife and children Leo and Gabrielle. Research in French archives then offered more insights into this individual and provided further information about the treatment of allied citizens in Indochina who were considered hostile by Japan. Furthermore, these documents shed light on the complicated relationship between French, Japanese and Allied citizens in Indochina. Gabriel Corvissiano was arrested, like most of his compatriots, soon after Japan entered World War II. However, he was treated differently because of his commercial activities: specifically, because he was a partner of the French citizen Charles Edouard Anthony in the operation of a mining concession. This was an economic sector of strategic interest to the Japanese, but since the French authorities sought to curb foreign influence in the economic sphere of the colony, the Japanese were only permitted to

58   Cornwell, Peter. & Herren-Oesch, M. (2019). Treatment of American National Archives Records of World War II Prisoners of War. (1.1.0). 10.5281/zenodo.3565392.
59   NARA *Record Group 389* available at https://aad.archives.gov/aad/series-description.jsp?s=644&popup=Y.

operate in cooperation with a French citizen.[60] Japanese intelligence recognized the opportunity presented by Anthony's circumstances, leading to the release of Corvissiano from detention on commercial grounds. However, the threat of re-internment remained. It appears that Corvissiano was reminded by the police of "benevolence" on the part of Japanese officials, intimidating him into relinquishing his commercial interest and making way for a Japanese partner in the business. He conceded his share of the concession in August 1942 and, together with his family, was repatriated to Lourenço Marques[61] (now Maputo, Mozambique) soon afterwards.

Result 3 of 14

Data resource: Asian Directories & Chronicles, Year 1941

Name: Anthony, C. E.

Locations: Haiphong

Anthony, C. E., comml. mgr., Societe des Verreries d'Extreme-Orient, Haiphong

view repository   (Asian Directories & Chronicles, Year 1941)

Result 4 of 14

Data resource: Asian Directories & Chronicles, Year 1935

Name: Anthony, C. E.

Locations: Haiphong

Anthony, C. E., comml. manager, Societe des d'Extreme-Orient, Haiphong

view repository   (Asian Directories & Chronicles, Year 1935)

Selected results of search for Anthony, C. via a website interface using Asia Directories scientific treatment data, showing transcribed text and tokenized values, as well as IIIF fragment defined by WADM annotation.

Charles Anthony can also be traced using the DPC treatment dataset for *Asia Directories* foreign resident listings, as described earlier and in the illustration above. We readily find eight appearances of an Anthony, C. E. in Haiphong between 1926 and 1945. These occurrences (as well as others sharing this surname), which can be viewed online at https://asia-directories.org), list the occupation of this Anthony consistently as Commercial Manager and they return Société des Verrieries des d'Extreme-Orient or Société des d'Extreme-Orient or Société Francaise des

---

60   Mining and agricultural concessions in Indochina had to be organized by a Franco-Japanese partnership. See Sébastien Verney. *L'Indochine sous Vichy: entre révolution nationale*, collaboration et identités nationales, 1940–1945 (Paris: Riveneuve éds., 2012), 205.

61   Court records in Archives National Pierrefitte-sur-Seine, CONTRE: Laubies, Anthony, Signature: Z/7/23.

Verreries d'Indo-Chine as employer. While these results alone are inadequate to establish a definitive Charles Anthony person entity, they are convincing support for Corvissiano's partner having resided in Haiphong prior to and after their business relationship.

Reconstructing key parts of this narrative in a few moments, via instances of persons occurring in unrelated documents, greatly increases the rate at which such hypotheses can be tested by scholars. The same scientific treatment data produced from these documents serves both this individual purpose and the identification of asymmetric movements of foreign residents in the Straits, compared with populations in mainland China, in the previous example. Here, the new DPCL data resources have provided insight into the politics of internment in Indochina perpetrated by the French authorities, as well as the challenges faced by Gabriel Corvissiano and his family. This would have been very laborious indeed to achieve using conventional methods. Corvissiano seems to be an exception: currently no other U.S. citizens have been identified who were allowed to pursue their profession or who were repatriated while these internment policies were operating. For his part, Charles Edouard Anthony realized that continued collaboration with a partner from an Allied nation was untenable, and precipitated Corvissiano's departure. The overwhelming Japanese military presence and increasing isolation of Indochina meant that Japan represented the only practical market for Anthony's business.

Concluding Remarks

A range of existing techniques and software applications have been applied to challenging historical documents to build DPC's scientific treatment data resources, with only modest investment in developing workflows for scholars to use at scale. Transferring fine-grain methods from the life sciences, multiple uses of these data resources have produced new insights into conditions in East Asia during the late 19th and early 20th centuries, where traditional progress based on histories of institutions had previously stagnated. Privileging practices has permitted the development of empirical data for history at a local level – global micro-history. Additional projects are planned, building on DPC's outputs, including collaborations with existing large-scale research investments such as Navigocorpus[62] and Swiss Elites.[63]

More important than this, however, is the re-usability and sustainability of digital outputs generated in the Divisive Power of Citizenship project. Scientific treatment data resources are general-purpose and technology-independent from the outset, eliminating the costs of separate preservation work or the need for cyclic redelivery when specific technologies become obsolete. Furthermore, significant new digitization by DCP at resolutions effective for digital preservation, as well as efforts to unite serial publications previously distributed in many individual institutions, are protected by the adoption of repository infrastructures and access standards which are supported by the physical and life sciences. Financial investment in other domains reduces otherwise prohibitive maintenance and long-term accessibility costs for SSH digital outcomes, which have until now undermined the implementation of the FAIR Principles.

While these strategies are now freely available to the broader research community, their immediate uptake and meaningful reduction in the loss and continuing vulnerability of research data is far from guaranteed. There is currently no evidence of research funding agencies preparing to articulate data preservation hazards more clearly, nor of providing practitioners with the means to take responsibility for long-term accessibility to interim research data, compared with published final outcomes. Until it is recognized that interim data produced *during the course of* research activities has value, there can be no framework for policies to protect it, and research investment will continue to evaporate. Like Laurie David's *An Inconvenient Truth*,[64] taking the step of estimating the volume of data currently being lost or the impact that this undoubtedly has on new research, is not a welcome task for funding agencies.

---

62   https://anr.portic.fr/en/sources/.
63   https://www.unil.ch/obelis/en/home.html.
64   *An Inconvenient Truth* was a 2006 American documentary film produced by Laurie David, an American environmental activist and writer, and directed by Davis Guggenheim, about former United States Vice President Al Gore's global warming campaign.

Bibliography

Agosti, Donat, and Egloff Willi. "Taxonomic information exchange and copyright: the Plazi approach." *BMC Research Notes 2*, no. 53 (2009). 10.1186/1756-0500-2-53.

An Inconvenient Truth. Directed by David Guggenheim. USA: Paramount Classics, 2006.

ARCADIA. Homepage. *ARCADIA*, 2023. https://www.arcadiafund.org.uk/.

Berisha, Blend, and Endrit Mëziu. "Big Data Analytics in Cloud Computing: An overview." Seminar paper, University of Pristina, 2021. 10.13140/RG.2.2.26606.95048.

Biodiversity Literature Repository. Homepage. *Biodiversity Literature Repository*, 2023. https://biolitrepo.org/.

CERN – *Conseil Européen pour la Recherche Nucléaire*. Homepage. *CERN*, 2023. https://home.cern/.

CETAF – Consortium of European Taxonomic Facilities. "CETAF to adopt IIIF standard for sharing images of natural history objects." *CETAF*, January 24, 2023. https://cetaf.org/elementor-7894/.

Cornwell, Peter. "Preserving Scientific Annotation Working Group Birds-of-a-Feather Meeting." *Zenodo*, November 7, 2018. 10.5281/zenodo.2633630.

Cornwell, Peter, and Madeleine Herren. "Treatment of American National Archives Records of World War II Prisoners of War (0326)." *Zenodo*, December 6, 2019. 10.5281/zenodo.3565392.

Cornwell, Peter, and Madeleine Herren. "Asian Directories: Foreign Residents Benchmark Dataset." *Zenodo*, August 30, 2019. https://doi.org/10.5281/zenodo.2580998.

DARIAH – Digital Research Infrastructure for the Arts and Humanities. Homepage. *DARIAH*, 2023. https://www.dariah.eu/.

DataCite. Homepage. *DataCite*, 2023. https://datacite.org/.

Fedora – Flexible Extensible Digital Object Repository Architecture. Homepage. *Fedora*, 2023. https://fedora.lyrasis.org.

Fedora Project. Homepage. *Fedora Project*, 2023. https://docs.fedoraproject.org/.

Futter, Christian, and Peter J. Cornwell. "Les ressortissants américains confinés à Mytho." *Zenodo*, July 9, 2022. 10.5281/zenodo.6813785.

Futter, Christian, and Peter J. Cornwell. "Les ressortissants canadiens confinés à Mytho." Zenodo, September 9, 2022. 10.5281/zenodo.7065068.

GBIF – Global Biodiversity Information Facility. Homepage. *GBIF*, 2023. https://www.gbif.org/.

Git. Homepage. *Git*, 2023. https://git-scm.org.

GitHub. Homepage. *GitHub*, 2023. https://github.com/.

GO FAIR. "Fair Principles." *GO FAIR*, 2016. https://www.go-fair.org/fair-principles/.

Herren, Madeleine. and Peter Cornwell. "Communication at the Dawn of the Golden Age: The Asia Directories and Chronicles. Combining historiography and data-science in a micro-global approach to foreign resident activity in East Asia during the 19th and early 20th centuries." Zenodo, March 17, 2020. 10.5281/zenodo.5579598.

Holm Nielsen, Lars. "InvenioRDM reaches major milestone - v6.0 released." *InvenioBlog*, August 5, 2021. https://inveniosoftware.org/blog/2021-08-05-inveniordm-lts/.

International Image Interoperability Framework. "IIIF: Access to the World's Images – Ghent 2015." *IIIF*, 2023. https://iiif.io/event/2015/ghent/.

International Image Interoperability Framework. "Consortium Members." *IIIF*, 2023. https://iiif.io/community/consortium/members/.

INSPIRE. Homepage. *INSPIRE*, 2023. https://inspirehep.net/.

Invenio. "History." *Invenio*, 2023. https://invenio.readthedocs.io/en/latest/community/history.html.

Invenio. "InvenioRDM." *Invenio*, 2023. https://inveniosoftware.org/products/rdm/.

Jefferies, Neil. "Oxford Common File Layout Specification." *OCFL*, October 7, 2022. https://ocfl.io/1.1/spec/.

JSON Schema. Homepage. *JSON Schema*, 2023. https://json-schema.org/.

Library Technology Guides. "Caltech selects TIND Library Management System." *Library Technology Guides*, July 13, 2015, Press Release. https://librarytechnology.org/pr/20852/caltech-selects-tind-library-management-system.

NARA. "World War II Prisoners of War Data File, 12/7/1941–11/19/1946." *NARA*, 2023. https://aad.archives.gov/aad/series-description.jsp?s=644&popup=Y.

National Gallery of Art. "International Image Interoperability Framework: Sharing Images of Global Cultural Heritage: Introducing Mirador." *National Gallery of Art*, May 5, 2015. https://www.nga.gov/audio-video/video/iiif/iiif-snydman-winget-6.html.

OBELIS. Homepage. *OBELIS*, 2023. https://www.unil.ch/obelis/en/home.html.

Open Archives Initiative. "Protocol for Metadata Harvesting." *Open Archives Initiative*, 2023. https://www.openarchives.org/pmh/.

ORCID. Homepage. *ORCID*, 2023. https://orcid.org.

Plazi. Homepage. *Plazi*, 2023. https://plazi.org/.

Popplow, Marcus. "Technology and technical knowledge in the debate about the 'great divergence'." *Artefact* 4 (2016): 275–85. doi: 10.4000/artefact.485.

Portic. "Sources." *Portic*, 2023. https://anr.portic.fr/en/sources/.

PyPI. "Ocflcore 0.1.0." *PyPI*, December 13, 2021. https://pypi.org/project/ocflcore/.

QGIS. Homepage. *QGIS*, 2023. https://www.qgis.org/.

RDA – Research Data Alliance. "RDA Research Data Repository Interoperability WG." *RDA*, 2023. https://www.rd-alliance.org/node/50279/case-statement.

Reinhardt, Anne. *Navigating Semi-Colonialism. Shipping, Sovereignty, and Nation-Building in China, 1860 – 1937*. Cambridge, Massachusetts: Harvard University Asia Center, 2018.

Stanford, Emma. "Fedora and Hydra/Samvera Camp at Oxford Sept 4-8 2017." *Bodleian Digital Library, A Bodleian Libraries blog*, July 6, 2017. https://blogs.bodleian.ox.ac.uk/digital/2017/07/06/fedora-and-hydrasamvera-camp-at-oxford-sept-4-8-2017.

Swiss Institute of Bioinformatics. "Swiss Institute of Bioinformatics Literature Services (SIBiLS)." *Swiss Institute of Bioinformatics*, 2023. https://www.expasy.org/.

Text Encoding Initiative (TEI). Homepage. *Text Encoding Initiative (TEI)*, 2023. https://tei-c.org/.

Verney, Sébastien. *L'Indochine sous Vichy: entre révolution nationale, collaboration et identités nationales, 1940 – 1945*. Paris: Riveneuve éds., 2012.

Vines, Timothy H. et al. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24, no. 1 (2014): 94 – 97.

W3. "Open Annotation Community Group." *W3*, 2023. https://www.w3.org/community/openannotation/.

W3C. "Web Annotation Data Model." *W3C*, February 23, 2017. https://www.w3.org/TR/annotation-model/.

W3C. "Open Annotation Data Model." *W3C*, February 8, 2013. http://www.openannotation.org/spec/core/.

Zenodo. "About Zenodo." *Zenodo*, 2023. https://about.zenodo.org/.

Zenodo: "Communities created and curated by Zenodo users: covid." *Zenodo*, 2023. https://zenodo.org/communities/?p=covid.
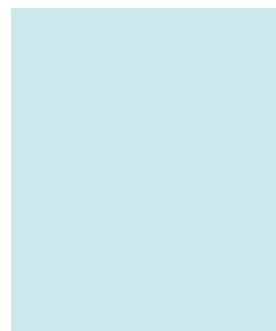
**Educating
Talents**
since 1460.

Europainstitut der Universität Basel
Institute for European Global Studies
Riehenstrasse 154
CH-4058 Basel

www.europa.unibas.ch