CHAPTER 5

Building Historical Knowledge Byte by Byte

Infrastructures and Data Management in Modern Scholarship

Jessica Parland-von Essen

Introduction

Historians are very good at source criticism, but in the digital era this requires good provenance data. Historians should also step up to the demand for transparency and open scholarship that comes with digital humanities. Research and knowledge has to be well documented and reliable. This means we need good data management, but also better and more integrated services and infrastructures.

Despite often exceptionally rich descriptive metadata in the cultural heritage sector, research life cycle data management is not easy and finding sources might be difficult due to questions of metadata formats or granularity of publication. The humanists' workflow and practices regarding use of sources is often hybrid and only partly digital.[1] In this chapter, I will analyse different digital data types and infrastructures from the point of view of a historian and discuss the needs of historical research and knowledge creation. Questions about data

How to cite this book chapter:
Parland-von Essen, J. (2020). Building historical knowledge byte by byte: Infrastructures and data management in modern scholarship. In M. Fridlund, M. Oiva, & P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history* (pp. 89–102). Helsinki: Helsinki University Press. https://doi.org/10.33134/HUP-5-5

management and information structures are important to solve, so that it is possible to formulate service needs and user stories for historical research data services. I will propose a model for planning research data management and data publication for historians. The chapter focuses on the Finnish research sector, but includes relevant international infrastructures and initiatives.

The Concept of FAIR Data

FAIR data was minted as a concept in an expert meeting among science data experts, and resulted in a seminal article on research data management published in 2016.[2] The concept, which was a more than needed completion to the Open Science, Open Access and Open Data rhetoric, won immediate approbation within the European Union and other data-aware stakeholders. It was obvious that open data or access was not by far enough to solve the issues with science reproducibility, let alone the efficiency goals of the Digital Single Market. Data cannot always be open and there were other, more technical hurdles, too. Data needed to better managed, and the money invested in research should not be wasted by sloppy planning. To make the most of our data, it has to be organised and taken well care of. Only then can we combine datasets and build digital knowledge by linking publications and data in sustainable and trustworthy ways.

In short, the FAIR principles state that data should be Findable, Accessible, Interoperable and Re-usable. It turns out that these fine words in practice result in very technical definitions. When going into details, we soon exceed the level most scholars in the humanities should have to be bothered with. We should simply have workflows and services that support these principles, but for that to happen, all stakeholders have to raise their awareness and understand what is necessary to accomplish regarding services and infrastructures.

Let's take a short look at the principles and how they could be translated into a relevant form for our purposes. F stands for *Findable*. What this actually means is machine-readable. The amount of data today is so immense that it is important that computers cannot only sort out data, but also act upon it and find what is really relevant. This means, for instance, not only that digitising text so that it is only in image form is not sufficient, but also that the content of text needs to be organised in more specific, semantic ways: it requires structured metadata and keywords, as well as common and persistent identifiers for concepts like persons or place names. Furthermore, the metadata has to be available for and utilised by different kinds of indexing and search tools.

A means *Accessible*. This, in practice, today means data that can be downloaded over the web, or at least the internet. Both machines and humans should be able to understand the information the data represents or contains, and it should not be transferred or changed in non-transparent or undocumented ways. I, as in *Interoperable*, is a tough one. It means you should be able to combine datasets and copy metadata smoothly, without losing any information.

This means you should comply with existing standards and formats. As research data management in many ways is in its infancy and the information systems are still largely insufficient or impractical, this is difficult. It is necessary to balance the needs of the research and serve the actual research use, which must be prioritised. Unfortunately, many researchers are inclined to think that their data is far more different and unique than it actually is or needs to be. Usually, it is possible to find *some* aspect of the data that somehow relates to something else, be it source, structure or some semantics of the content. As people tend to understand how much effort they have put into their own work and development, it is too easy to underestimate the value of other people's work. The *not invented here* syndrome[3] can easily trump real creative openings and slow down research. Particularly in the life sciences, there have been many important insights and tools developed (bioinformatics might be the oldest domain-specific field within research data management). We should copy as much and as fast as we can from other, successful domains.

R, which is *Re-usable* data, means that it has a functioning licence or rights statement, but also that it has been thoroughly documented so that another researcher, or the composer of the dataset in 10 years for that matter, can take a dataset and use it again. Often, researchers spend up to 80% of their time creating or cleaning their data.[4] Therefore, careless documentation can be considered an inexcusable waste of resources and time.

The utmost goal, besides efficiency, is of course trustworthy, high quality research. The digital environment has the unfortunate quality of being simultaneously dynamic and unreliable. Links, even in scientific publications, tend to break.[5] This phenomenon is called link rot. Similarly, the content behind the link might change in a devious, unnoticeable way, which is called content drift. To address this problem, one of the main building blocks of FAIR data are *persistent identifiers*. Above, I mentioned identifiers for different kinds of concepts, which makes it easy to trace and link information. Researchers might have their own identifiers in the form of an ORCID, which is personal, unique and resists changes in name form or affiliation, and makes it possible to differentiate people with the same or similar names. Correspondingly, the datasets and articles should have their own identifiers, a URN or a DOI, which makes citing clear and unambiguous. The point is then the persistence; namely, the sustainability of this identifier. This means that we need platforms and services that provide and manage them on a long-term basis. This has a direct connection to the importance of infrastructure, which I will address later in this text.

To a historian, it is obvious that one has to address problems of sustainability in the long-term perspective, as well as that the sources need to be well documented. Are there other means for evaluating the trustworthiness or suitability of the data for our needs? Or to ensure that the data are authentic and have maintained their integrity? We need to know who said what, where and when. *Simultaneously, we also need to accept that our own research outputs should meet these requirements*.

The example of citations, the ultimate goals and tests for the data, demonstrates well the problem of sustainability. We should ask ourselves how can I cite (link to) my (digital) source in a persistent and unambiguous way and how can someone else cite the data I have created? There are recommendations for this, but they are not obviously sufficient or easy to implement. The national Finnish guideline for citing research data offers principles for citing a dataset, but how to cite more dynamic resources and what to do[6] when the resource does not provide identifiers or possibilities to download or save (partial) snapshots? Or even if the researchers manage to download the needed data, where do they archive it conveniently? The questions of data management during research are inescapable for all these practical, technical reasons. However, data management is even more complex for historians, because of questions about personal data regulation, ethical issues and copyright.

The Historian's Data Life Cycle

In Finland, the government and major research funders have promptly adopted the Open Science ideology, and research data was included in the policies at an early state.[7] There has been quite extensive work done on a national level regarding services, formats and recommendations. In parallel, there has been an effort for interoperability and digital preservation within the cultural heritage sector. This has produced services like the search portal Finna.fi and the national preservation services.[8] These and their future development are of course both important from a historian's point of view. Still, the situation for research data is quite different, since research data does not come with a clear legislation, accountability and centuries-old tradition of long-term or even short-term management. Responsibilities are often unclear when it comes to both rights and costs. In the humanities, researchers are used to expecting free or subsidised services when it comes to sources and information management. On the other hand, the research outputs are clearly considered to be the property of the researcher, at least concerning copyright. The work within humanities is considered creative and personal and thus often falls under intellectual property rights legislation.

The problem is, of course, that ownership is not a simple concept when it comes to digital resources. There are many kinds of rights and responsibilities entailed in 'owning': who has the right to access, copy, use, give access, agree on use, alter or destroy a dataset? Who has the responsibility to keep the platforms running, create metadata, plan for migrations, manage access for the next decades and curate the metadata or data if errors are found? It sometimes seems that some believe that the researcher herself should have all the rights with no responsibilities, even after the research has ended. This obviously does not work. There has to be an agreement and a balance in responsibilities and rights management. The researcher might have to give up some of the control

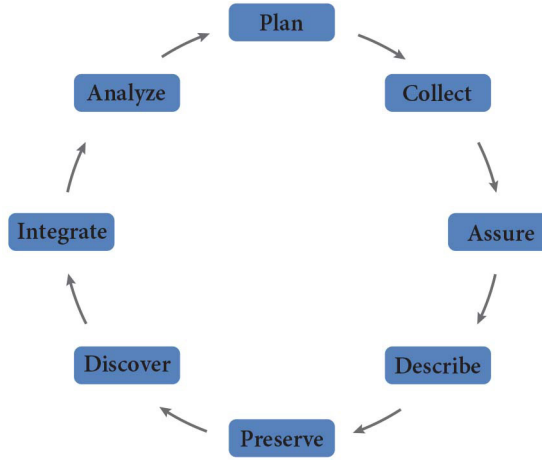

Figure 5.1: The DataONE data life cycle. Source: Author.

of her data in return for someone taking care of it. This calls for trust from both parties and concordance on common interest and explicit agreements. This is usually not a problem, but problems tend to arise from insufficient research data management planning. The agreeing would best be done in advance, preferably not by dictates from one party or the other, but by joint interests which should be easy to identify. Since the historian rarely makes up the data, but refines existing digital or non-digital data, there are usually concerns that need to be taken into account already when the data is created. Therefore, the data management life cycle always starts with planning.

There are different interpretations of the research data life cycle, but generally they tend to be variations of models that reflect the traditional way of understanding how the research process works in theory (see Figure 5.1). The idea is that there is always a project and one or several funders. Although often presented as a circular, never-ending process, one premise seems to be linear progression of the research process, as well as of science and knowledge building. This is, as any historian or other researcher knows, obviously a construct that nicely resonates with the way in which scientific publication traditionally works, with outputs that are corresponding, constructed narrations about the research process. The reality is much more chaotic and unorganised, which any data librarian will also willingly admit. The traditional publishing comprises snapshots, reports frozen in time, documenting what has been done, for dissemination and future reference. Still, these knowledge bytes are cumbersome, ambiguous and digitally discrete from the sources.

Thus, the single 'byte' of new knowledge has actually been quite open for future interpretation, often difficult to spot and point to. Even though the novelty might be a new interpretation or insight, there might also be included other new information or 'factoids', all of which become buried within an

extensive narration impenetrable for computers. Much of the information now being digital, there might be an opportunity to critically assess how we communicate our knowledge and are open to transformation within scientific publishing. Often, digitisation has meant diversification as well as convergence.[9] When we now bring data into the world of publication, there are immense possibilities for opening the whole process, enhancing documentation and sharing knowledge in new ways.

The historian has to decide upon how many of the sources can or should be linked to, in other words how many should be digitised, if the sources aren't digital and how digitisation should be achieved. Or perhaps the links are all external, linking to existing trustworthy digital sources? Data collection and creation is more complex in digital humanities than in traditional humanities research, since questions of documenting provenance and deciding on data and metadata formats will affect the research in profound ways. There are some cases where established standards exist, like TEI (xml format by the Text Encoding Initiative) for encoding text. But TEI in itself will not solve problems of interoperability on a deeper semantic level. It would, for instance, always be advisable to use good external references as identifiers for all concepts whenever possible. Also, the plan for publication might set limits to what the researchers should do, since the platform they choose might have some bearing on the formats, metadata and granularity of the publication. If the researchers use other people's digital resources (OPEDAS or Other People's Existing Data and Services, as named by a leading FAIR data expert Barend Mons[10]), they obviously need to find out extensive information about them, not only the technical and historical provenance, but also about how the data is structured and coded. Often, a historian uses OPEDAS created not by researchers, but by heritage institutions. As the use context changes, the data provider institutions generally do not have readymade generic solutions for managing and publishing research data, especially when it is produced by outsiders.

One of the unfortunate traits of the traditional data life cycle model is that publication turns up as a distinct step in a specific and late stage of the process. This hides the fact that the most efficient and impactful way of doing research might be doing it transparently *all the way*. Since this both forces the researcher to implement some type of data management and opens up for collaboration and spotting quality issues at early stages, this can accelerate the work and enhance the quality of the research. After publishing raw versions of data, unforeseen help can turn up, when colleagues become aware of what the researcher is doing. Close collaborations have not always been an option in historical research, which carries the heavy burden of romantic lonely genius syndrome, but luckily times are changing. Stealing other people's ideas and data is not the first thing most researchers think of. Rather, by publishing raw data, the researchers can get their work registered at an early stage, instead of waiting for the final peer review. Better collaborating and coordinating than working in silence.

Version control is the next crucial aspect of the data cycle. If you ask an archivist, they would probably want to save every version of everything. Even worse, this might mean not just saving the information you need to recreate the needed version of a dataset, but saving complete copies of each version, independent of all redundancy that would create. Version control is generally not that well developed in traditional archives. However, every version that is published needs metadata and, preferably, a persistent identifier. But this does not mean that the researchers have to save everything, every single byte. The researchers simply have to be sure that the dataset can be presented in an exactly identical form when asked for at a later point in time. In case somebody made a citation or important conclusion based on it, it should be possible to reconstruct what has happened. It is very important to be clear about it, if this is something the researchers do *not* commit to, when they publish data.

Managing research data is not the same thing as archiving it, and handling digital data requires a somewhat different approach. Here, storage and data management are relevant components building trustworthiness of the documentation. Citation is one of the main functions of persistent identifiers in research. The researchers should be mindful creating them though, since every persistent identifier is a commitment to manage the dataset or at least its metadata forever. It will cost somebody a substantial amount of effort and work. And even if the dataset is deleted, a tombstone page should be maintained. Here, the well-managed research infrastructures and data services come into the picture as essential supporters of research.

Generally, one could consider there to be three different types of datasets that are relevant for historians (see Figure 5.2). First, there is the master data produced and often published by government institutions, like the cultural heritage data. Unfortunately, it is not always well versioned or documented (red in Figure 5.2). It could be data of any kind for any use, but it might be relevant for a historical research question due to a long time series or for some other reason. Second, there are generic research datasets, which are produced by researchers for scientific use (green). Here, you find datasets like corpuses or some of the surveys published by the Finnish data archive. Much data of this kind can also be found, for instance, with the National Institute of Health or other domain-specific research institutes or government bodies. These datasets are validated and often cumulative. The third type of research data is a research output, created to underpin a specific study or article (blue). These data need to be saved, albeit the interest for reuse might be minute, for the simple reasons of reproducibility of the research and merit for the creator.

The historian often finds her digital sources within the first or second category of data. But as she proceeds with her work, the question of publishing second- or third-type data becomes increasingly pressing. Now, there is no single clear path to publishing this kind of data, which is often a derivate of cultural heritage data. Additionally, researchers within the humanities many times deal with sensitive data or data under copyright, which makes storing

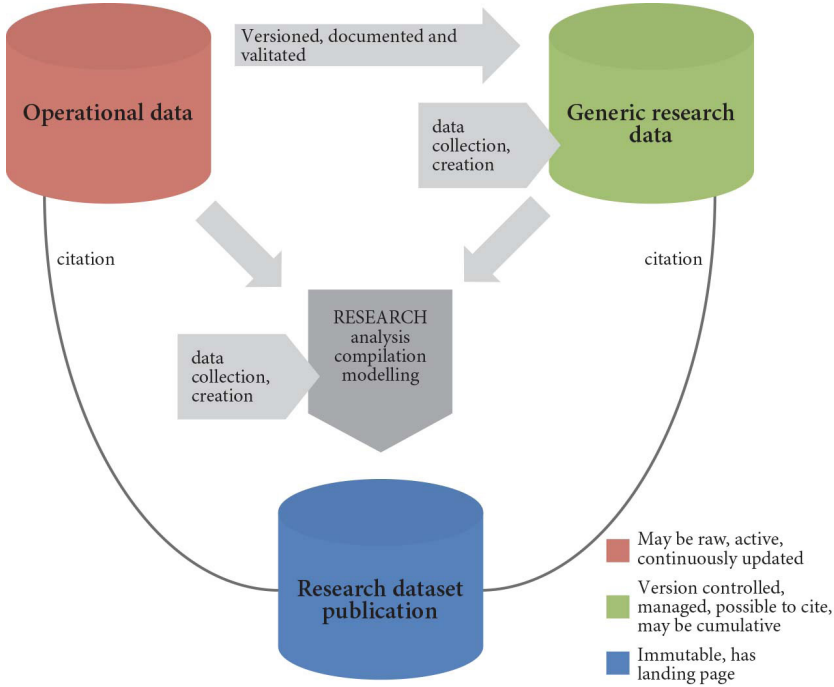

Figure 5.2: Main types of data used by historians and how those are interrelated. Source: Author.

and publishing even more difficult. I will discuss the options later in this chapter when discussing research infrastructures.

There is often more to a digital humanities research outcome than just a dataset and a result explained in a narrative form. It needs to be pointed out that the historian often handles a double narration: one of the research process and then another, which is the actual new knowledge. This is the normal situation when carrying out qualitative research or being unable to present or refer to the actual tools and methods used. However, when using computers and computational methods, the process and outcomes like dynamic databases or visualisations could and should also be included in the outputs, in addition to information about the sources or actual data. For this, the existing solutions are few and the methodology is very thin. Preservation of databases has developed somewhat, but documentation and preservation of dynamic user interfaces and other kinds of complex code is still in its cradle. It is well known that they need an extensive amount of curation to be kept usable for more than some years. This means that they are both risky and costly to preserve. Still, some effort to save these is better than just abandoning digital projects at the end of project funding. The problem is usually to find the party willing to take the responsibility. Therefore, this is also one thing that best would be solved at the point of planning the research.

Source Criticism and Research Assessment

Assessing digital sources requires a substantial amount of metadata. I need to discuss this theme more closely to explain why and how data management planning and infrastructures are relevant not only for creating FAIR data, but also for carrying out high quality research in history in a digital environment.

A digital document does not have an 'original' copy. Instead, it is recreated every time the source is rendered from a digital file consisting of zeroes and ones. Everything is just copy, while the analogue versions, which are the ones we can perceive with our senses on the screen or in our ears, are generated by software and hardware that have a decisive effect on what they actually apprehend. The calibration of the screen or the sampling frequency of an audio file might affect how one interprets what is represented or real. In cases where a physical original exists, one can always check it, but if the source is born-digital, this becomes impossible. Therefore, there is a need for technical metadata.

The best way to evaluate the trustworthiness of a digital source, as is commonplace for the historian, is to check its provenance. In practice, the researchers need to assess the organisation or person who has delivered the source. Can they show documentation about the technical and administrative life cycle of the source? Do they comply with the Open Archival Information System (OAIS) standard or do they have other certificates for digital preservation?[11] Do they use and manage persistent identifiers that are globally unique and persistent? Can they present extensive metadata, including checksums? The checksums are important digital seals for calculating the integrity of files, but they do not work across file formats, which is why the researchers need to have a good trail of documentation and management of persistent identifiers. The formats might have changed during the life cycle of the data. What else has happened in terms of migrations, curation, cleaning and enhancing the data? Is everything convincingly documented?

There are several kinds of metadata. To be able to represent a digital source in a similar or corresponding way we need technical and structural metadata that helps one choose the right tools and understand possible offset. We also need administrative metadata that informs about the rights and responsibilities attached to the data. Furthermore, we need descriptive metadata, which helps with finding and organising the data, as well as with the usual historical source criticism around what, who, when, why and other contextual information. This is the part of information that is most threatened in research data, due to reasons of personal data. Data archives often prefer anonymised data, which means crucial historical information is permanently lost from the historian's point of view. This is also the reason why the current research data archives do not provide sufficient services for many historians. The organisations that do this best are institutions like the Swedish and Finnish literary societies, which have a profound understanding of the importance of the personal and unique as part of the greater whole and of the research processes within cultural studies and history.

It is important to understand the ephemeral nature of digital information, not only when it comes to the historian's own sources, but also concerning working with data. If research is to be possible to repeat, the digital operations undertaken should be well documented. Code should be documented and saved and versions of the dataset have to be managed. Not everything has to be saved, but one should consider versioning and documentation when significant changes are made.[12] Conversions, cleaning and mapping need to be accounted for, since they may affect the outcome of the research. And as technologies become obsolete over time, all types of metadata are necessary. Otherwise preservation will not be possible.[13] This part of the data management should be planned together with data librarians and professional data stewards.

Infrastructure and Services

Reliable and good quality research craves good citations and linking. The historian's digital sources can be found in cultural heritage institutions or in many other places that sometimes, but not always, offer possibilities to create FAIR data by pointing to the sources in exact and sustainable ways. Often, the researcher needs to clean and organise the data, which in turn creates a new dataset.

According to the European Commission, research infrastructures (RIs) are facilities, resources and services used by the science community to conduct research and foster innovation. The Finnish Academy is lengthier in its definition:[14]

> Research infrastructures refer to a reserve of instruments, equipment, information networks, databases, materials and services enabling research at various stages. Research infrastructures may be based at a single location (single-sited), scattered across several sites (distributed), or provided via a virtual platform (virtual). They can also form mutually complementary wholes and networks. Europe hosts several large-scale research infrastructures that are open to collaborative use across national boundaries.

The Open Science and Research Initiative report addressed RIs.[15] This report distinguished between services, data and equipment. This classification has also been implemented in the national Research Infrastructure catalogue, which provides persistent identifiers for these (https://research.fi/).[16] Many infrastructures provide two or three of these types of resources. The national strategy for RIs[17] demonstrates that we have advanced infrastructures for linguistics, register research and social sciences. The national consortium for supplying digital publications for the research libraries within all domains is also, for some reason, considered a humanities and social sciences infrastructure.

The cultural heritage sector is left out, except for the shared search portal Finna, which serves the public as well as the research community at large when it comes to traditional research publications (namely, articles and monographs). This means the search portal aggregates some relevant research data for historians, like individual photographs or archival collections, research dataset metadata from the Data archive (a patchwork with very few internal links and of highly varied granularity) and research literature from all fields. The European cultural heritage portal Europeana does the same, except leaving out the literature and focusing on the traditional but digitised sources.

The main problem is, besides missing sufficient persistent identifier management, the lacking information structures. The digital objects vary in size from single photographs to archival collections and corpuses with almost nonexisting descriptive metadata from a historian's point of view. Saying this, I do not want to belittle the enormous and important work that has been done to bring all this metadata together. It has been an extremely valuable effort, with thorough implications for the cultural heritage sector in Finland, which has taken huge steps towards openness and digitisation. However, for research we still require better representations of the sources and their internal relations. Important digital sources are omitted, including databases provided by the institutions themselves, not to mention historical research databases elsewhere, whose producers often face great difficulties getting hold of sustainable funding or sufficient data management for their digital research outputs. The cataloguing of these resources, documentation and linking datasets derived from cultural heritage data in general is today left to the researcher, who generally has few possibilities to maintain these after the funding ends. Today, the historian most often has to be content with publishing discrete research datasets as simple files, which have weak and only human-readable links to other digital resources. Also, the reuse value is less than it probably would have to be, due to this approach and meager machine-readable semantics.

Both the Language Bank of Finland and the data archive have juridical mandates to store this kind of data, but the researcher has an extensive responsibility too. The slightest flaw in consents or rights questions easily becomes an insurmountable hurdle for archiving or sharing the data. There are also reasons to question whether this kind of publishing is the one and only, or whether there could be more suitable platforms or structures than the currently available solutions.

Digital media are not only unstable and diverse, but they are also often more disposed for interactivity and a dynamic communication that happens in dialogue, even co-creation with the readers/users.[18] In fact, it might be a mistake not to consider this kind of publishing and knowledge creation in a research domain that is so relevant and open to popularisation and popular culture. Different kinds of map and wiki applications can be used for sharing historical knowledge. Wikis are especially suitable due to their very transparent and clear version management. They also enable very good structuring and linking of

data.[19] In fact, the wiki technology combined with careful data management would offer an almost out-of-the-box solution for FAIR data.

The historian needs to carefully plan her data management. Questions of personal data, consent and copyright need to be addressed at an early stage before even starting the research. This does not mean that one has to decide on every detail or stick to the plan whatever happens. In fact, the opposite is often true: the plans have to be modified or redone, when new issues arise. The research process in digital humanities is often iterative, oscillating between qualitative and quantitative methods, and research questions sometimes have to be adjusted or revised.

From the very beginning, it is important to plan for managing data files, backups and versions. Also consider the types of data that will be included and analyse the need for documentation needed for citations and reproducibility. It is not necessarily a good idea to get a resolvable persistent identifier for every single data object. Instead, one should be pragmatic and consider the dataset as a part of the surrounding information universe and try to create meaningful, machine-readable and sustainable relations to that universe. Do not produce new data objects where you can reliably point to external ones. Also, one should be mindful about the granularity: Which are meaningful entities to make findable and for which to create metadata?

When it comes to infrastructures, we have to operate with what we have got, but historians could also give valuable input in creating a meaningful larger network of digital historical knowledge by engaging even more in questions of common or interoperable infrastructures. There are large infrastructure initiatives like DARIAH-EU, CLARIN-ERIC, Europeana and the European Open Science Cloud (EOSC), but there is still not a suitable solution that would serve historians well in publishing and linking their research outputs. It is essential that historians discuss these questions with other stakeholders, the cultural heritage institutions, the scientific libraries and their own research institutions and funders to find sustainable solutions and drive infrastructure development in directions that serve knowledge creation, not only as separate projects, but as a linked network of information.

Notes

[1] Antonijevic & Stern Cahoy 2018.
[2] FORCE11; Wilkinson et al. 2016.
[3] Not invented here 2018.
[4] Data science report 2016.
[5] Klein et al. 2014; Jones et al. 2016.
[6] Finnish Committee for Research Data 2018; Research Data Alliance 2015.
[7] Parland-von Essen 2017; see also openscience.fi.
[8] See Finna.fi, kdk.fi and digitalpreservation.fi.

[9] Anderson 2007; Manovich 2013.
[10] See Mons 2018.
[11] See, e.g., the DCP online guide on OAIS: Lavoie 2014 and the standard **ISO 16363:2012**.
[12] Language Bank of Finland.
[13] PREMIS preservation metadata.
[14] Academy of Finland 2018b.
[15] Avoimuuden politiikat tutkimusinfrastruktuureissa: Selvitys 2015.
[16] RIs, https://research.fi/.
[17] Academy of Finland 2018a.
[18] Salgado 2009; Nygren 2013; Marttila 2018; Viinikkala et al. 2016.
[19] See, e.g., Wikisources, Wikimedia, Wikidata and Tieteen termipankki. See also Wikidocumentaries and Wikimaps.

References

Academy of Finland (2018a). *Finland's strategy and roadmap for research infrastructures 2014–2020*. Interim report. Retrieved from http://www.aka.fi/globalassets/tiedostot/aka_infra_tiekartta_raportti_en_030518.pdf

Academy of Finland (2018b). *Research infrastructures*. Retrieved from http://www.aka.fi/en/research-and-science-policy/research-infrastructures/

Anderson, C. (2007). *The long tail*. Random House, London.

Antonijevic, S., & **Stern Cahoy, E.** (2018). Researcher as bricoleur: contextualizing humanists' digital workflows. *Digital Humanities Quarterly, 12*(3). Retrieved from http://www.digitalhumanities.org/dhq/vol/12/3/000399/000399.html

Avoimuuden politiikat tutkimusinfrastruktuureissa: Selvitys. (2015). *Avoin tiede ja tutkimus -hanke, Avoimuuden politiikat -työryhmä*. Retrieved from http://urn.fi/URN:NBN:fi-fe2016122731714

Data science report. (2016). *Crowdflower*. Retrieved from http://visit.crowdflower.com/r/416-ZBE-142/images/

European Commission. *About research infrastructures. What are research infra structures?* Retrieved from https://ec.europa.eu/research/infrastructures/index.cfm?pg=about

Finnish Committee for Research Data. (2018). *Tracing data: data citation roadmap for Finland*. Retrieved from http://urn.fi/URN:NBN:fi-fe201804106446

FORCE11. *The FAIR data principles*. Retrieved 29 September 2018 from http://www.force11.org/group/fairgroup/fairprinciples

ISO. 16363:2012: *space data and information transfer systems. Audit and certification of trustworthy digital repositories*. Retrieved from http://www.iso.org/standard/56510.html

Jones, S., Van de Sompel, H., Shankar, H., Klein, M., Tobin, R., & **Grover, C.** (2016). Scholarly context adrift: three out of four URI references lead to

changed content. *PLoS One, 12*(1), e0171057. DOI: https://doi.org/10.1371/journal.pone.0167475

Klein, M., Van de Sompel, H., Sanderson, R., Shankar, H., Balakireva, L., Zhou, K., & **Tobin, R.** (2014). Scholarly context not found: one in five articles suffers from reference rot. *PloS One, 9*(12), e115253. DOI: https://doi.org/10.1371/journal.pone.0115253

Language Bank of Finland. *Life cycle and metadata model of language resources*. Retrieved from https://www.kielipankki.fi/support/life-cycle-and-metadata-model-of-language-resources/

Lavoie, B. (2014). *The Open Archival Information System (OAIS) reference model: introductory guide*, 2nd edn. DPC Technology Watch Report 14-02. DOI: https://doi.org/doi.org/10.7207/twr14-02

Manovich, L. (2013). *Software takes command*. New York, NY: Bloomsbury Academic.

Marttila, S. (2018): *Infrastructuring for cultural commons*. Espoo: Aalto University Series, doctoral dissertations.

Mons, B. (2018). *Data stewardship for open science: implementing FAIR principles*. New York, NY: Chapman and Hall/CRC.

Nygren, T. (2013). Digitala material och verktyg: möjligheter och problem utifrån exemplet spatial history. *Historisk Tidskrift, 133*(3), 474–482.

Parland-von Essen, J. (2017). Från open access till open science. Framväxten av öppen forskning och vetenskap. *NORDICOM-INFORMATION, 39*(1), 97–103. Retrieved from http://nordicom.gu.se/sites/default/files/kapitel-pdf/von_essen_97-103.pd

PREMIS preservation metadata. Retrieved from http://www.loc.gov/standards/premis/

Research Data Alliance. (2015). *Data citation of evolving data*. Recommendations of the Working Group on Data Citation. Retrieved from https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf

Salgado. M. (2009). *Designing for an open museum: an exploration of content creation and sharing through interactive pieces*. Taideteollisen korkeakoulun julkaisusarja A 98.

Viinikkala, L., Yli-Seppälä, L., Heimo O. I., Helle, S., Härkänen, L., Jokela, S., Järvenpää, L., Korkalainen, T., Latvala, J., Pääkylä, J., Seppälä, K., Mäkilä, T., & **Lehtonen, T.** (2016). *Reformation representation. Mixed reality narratives in communicating tangible and intangible heritage*. DIHA & NODEM Special Session at 22nd International Conference on Virtual Systems and Multimedia VSMM, Kuala Lumpur.

Wikipedia. (2018). *Not invented here*. Retrieved from https://en.wikipedia.org/wiki/Not_invented_here

Wilkinson, M., Dumontier, M., Aalsbersberg, J. J., Hoekstra, H. E., & **Boyer, D. M.** (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. DOI: https://doi.org/10.1038/sdata.2016.1