

Libraries as e-infrastructure

Libraries have served as education and research infrastructures for centuries. In this paper, we will describe major opportunities and future challenges in the context of digital research and the »e-infrastructure« that are required for e-science. We will provide examples of current involvements and focus on the importance of cooperation at local, international, specifically European, and global scale.

Bibliotheken fungieren seit Jahrhunderten als Bildungs- und Forschungsinfrastrukturen. In dem vorliegenden Aufsatz werden die Chancen und Herausforderungen von digitalen Forschungs-umgebungen und von – für die sogenannte E-Science benötigten – E-Infrastrukturen erörtert. Es werden aktuelle Beispiele beschrieben; außerdem wird aufgezeigt, wie wichtig Kooperation auf lokaler, internationaler, speziell europäischer Ebene in diesem Zusammenhang ist.

LIBRARIES AS INFRASTRUCTURES

For centuries, libraries were a major, if not the main research infrastructure of academic institutions. They started off by holding the manuscripts and prints of researchers working at the institution, in times when reproduction of scholarly work was the exception and scholars had to travel around the world to gain insights into the works of other scholars. When the reproduction of scholarly works became easier, libraries were able to collect a large segment of the world's knowledge and make it accessible to researchers and students. Libraries' estates were usually established at the heart of the campus to perform their organizational function for the *circulation of knowledge* and serve as a *sanctuary for study*, where researchers and students could be among themselves and could receive advice by librarians – in many cases scholars themselves.

In parallel, experimental research became a major paradigm and laboratories containing a scientific apparatus became a major part of the institutional infrastructure. And laboratories started to produce knowledge resources that were usually not kept in libraries, namely research data and artifacts that underpin research findings. Libraries and laboratories – text and data – coexisted in a highly entangled form as research infrastructure partners for more than 300 years. Today, three rapid and radical developments bring libraries as infrastructures to a whole new level. First, digital knowledge resources are largely location independent. Second, and relatedly, research has become collaborative and distributed. Third, and most significantly for our question about the role of libraries as research infrastructures, data and the software used

to process it – forming compound objects representing virtualized experimental artifacts – became primary research outputs themselves. As text, data and software become more and more integrated, the resulting challenge for research infrastructure is how to sustain these *new* research objects.

Text expressing the researcher's narrative of ideas and methods was long time the sole authoritative record of research. Libraries have been keeping and providing access to this record for the society independently of changing publishing mechanisms. Since there are now new forms of compound knowledge objects that need to be kept as authoritative records of research, the question is how libraries, laboratories and computing centers can work together to maintain a record of research that can be reliably accessed now and by future generations.

Libraries and e-science

The rules of the scientific and educational system have changed tremendously with the use of information and communication technologies. Huge amounts of data are produced and can be immediately made available, interpreted, processed, enriched, stored and preserved. The old paradigm that access to research output is slow, difficult and expensive in order to be high quality is no longer valid. Traditional mechanisms for guaranteeing quality, such as peer review, have shown not to be 100% reliable and seriously slow down the review process. Expensive licenses have made access to research output hardly affordable for most research institutes.¹ Furthermore, copyright systems lack flexibility to allow for text and data mining.²

As an answer to the need for quick, easy, affordable and permanent access to research output, libraries have built digital repositories. A repository brings together all scientific output of an institution or a project. Libraries are widely recognized as a superior source of quality content, but they need to make more effort to increase the visibility of the content stored in these repositories.³ According to several studies, large amounts of papers (10–90%, depending on the field) published in academic journals remain uncited. Libraries can contribute to a more efficient and transparent scientific ecosystem in the e-science age. Interoperability standards, metadata enrichment, linked data,



Wolfram Horstmann

Foto: Bodleian Libraries



Wouter Schallier

Foto: privat



Jarkko Siren

Foto: privat



Carlos Morais-Pires

Foto: privat

and convergence of metadata schemes will give high quality scientific output more visibility. Libraries also need to aim at a full integration of formal publications (books, papers) with other content types such as grey literature, research data, software, audio, video, learning objects, etc. Finally, repositories give governments, funding agencies, and research institutes insight in the impact of the research that they support.

Since preservation of research output is no longer limited to institutional and format related boundaries, preservation becomes more complex. On the other hand, it is also an opportunity for libraries to organize preservation as a collaborative, global effort. The care for educational and scientific information as a public good represents also challenges for governments and policymakers. The emerging *compound knowledge objects* produced in collaborative research activities require a diverse set of services beyond the basic remit of storage; they should include easy to use services for deposit, registration, quality control, discovery, and access. These are supplemented with information-age infrastructure elements, such as semantic standards, specialist query and visualization tools, preservation services and elements which sustain critical characteristics of the repository materials: their integrity, authenticity, usability, and their ability to be understood and discovered.

Libraries in e-infrastructures

To derive greatest benefit from research data and any other form of research output, it is fundamental that library services for e-science are connected to state-of-the-art information and communication infrastructures, also termed e-infrastructures. These infrastructures include high-performance computing resources, fast networks, as well as information storage, access and management structures. Thanks to a long history of co-operation, libraries are well suited to develop digital information infrastructures as a collaborative effort. Recent examples of such innovative efforts involving big consortia include OpenAIRE (Europe)⁴, SHARE (USA)⁵ and COAR (global)⁶. Examples of services include coordinated advocacy and support (e.g. OpenAIRE National Open Access Desks or NOADs)⁷, information aggregation services building on institutional repositories, reporting services for research funders and institutions, and integration of all research outputs in »enhanced publications«, »executable papers« and finally through researcher workflows.⁸ The concept of Virtual Research Environment has been proposed as a working environment – for all sorts of scientific disciplines – that integrates all these elements and connects them with the underlying e-infrastructure.⁹

Research libraries and data centers are both immersed in the transition imposed on them by the adoption of e-science practices by the communities they serve.¹⁰ The complementary role they are adopting as providers of e-infrastructure services were described by the ODE project¹¹. The services provided by libraries and data centers must necessarily be aligned to provide the integrated data and text products as well as comprehensive workflows that can best support e-science research practices.¹² The study on Authentication and Authorisation Infrastructures (AAI) in research conducted jointly by LIBER (Association of European Research Libraries) and TERENA (Trans-European Research and Education Networking Association) is an example of how libraries and data centers can collaborate in developing common services to support e-science. It includes case studies that show how inter-institutional collaboration can be improved through the libraries' involvement in e-infrastructures.¹³ More generally, libraries have a significant potential to provide information services for collaborative science.¹⁴ The DataONE¹⁵ project and infrastructure also illustrates how libraries collaborate and provide services in linking data with publications as well as support for research data management. The FORCE11 community initiative¹⁶ involves many librarians in their efforts to improve scholarly communication, including enhanced publications as well as citation of research data. In this context, libraries execute the institutional implementation of global approaches for providing unambiguous research information, e.g. ORCID¹⁷ or FundRef¹⁸ for authors and academic institutions. And, of course, libraries are getting heavily involved in research data management.¹⁹

It is crucial for libraries to be involved in the development of infrastructures that ensure new ways of using scientific information,²⁰ a task that may require new partnerships. In e-science this includes the creation of machine-readable scientific records and text and data mining tools. Libraries need to participate in the current debate on legal reforms relating to these technologies (a summary of the discussions as well as proposals for reform are described in the report of the Text and Data Mining Expert Group²¹).²²

Libraries as sustainable hosts

In today's quickly evolving research world, libraries provide a sustainable framework for specialized services. ArXiv²³, probably the world's most renowned open access repository, is operated by Cornell University Library. PubMed²⁴, the world's most authoritative bibliography in the medical and life sciences is operated by the National Library of Medicine. And DataCite²⁵,

the world's most important service for providing persistent addresses for research data in the Internet is managed by the Technical Information Library of Germany in collaboration with many libraries such as the British Library and the California Digital Library as well as many research institutions around the world.

— New skills for librarians

Libraries have transformed their skillset over the last decades. In order to adapt to the researcher's new requirements, libraries had to hire business analysts and staff with academic background for digital scholarship support. Digital library systems require highly trained administrators and developers, and repositories require metadata as well as copyright specialists.

Since research data and software have become primary research assets that often require guarantees for permanent access, libraries can provide a safe harbor for digital research objects in a dynamic environment of mobile researchers, volatile repository content, transient products and short-lived standards. Libraries now need to tackle the challenge of making data and software reliably accessible and re-usable. This requires a transformational approach to library services and development of the new skills. Tasks such as the curation and stewardship for new research objects – data and software – will imply a profound revision of library and information science curricula, certificates and trainings, direct involvement in research projects, as well as learning on the job. Librarians will not become experts in data analytics, which is evolving as its own discipline. But they can become stewards who provide a sustainable basis for data scientists to work on.

— Local cooperation

The new roles of libraries in e-infrastructures have significant implications for the cooperation across the campus or the research institution. New forms of cooperation with researchers are emerging: one-to-one support and copyright advice for depositing publications, but also data consultancy in information intensive research projects. Interfaces to the financial and administrative systems of the research institutes need to be made in order to reliably link publications and data to research projects. And in all instances libraries need to closely align their activities with the computing services of the research institute to enable a seamless operation of services. Thus, virtual teams across libraries, computing services and research offices are being set up to tackle new challenges such as Open Access publishing and Research Data Management.

— Global cooperation

The grand challenges of the 21st century transcend borders, and science will be increasingly global. Data-driven science will require extensive global collaborations and researchers on each continent are striving for a leading role in the world's production of knowledge. Research data itself is global and the key issues to consider are:²⁶

1. How data can be networked
2. How to envision and set up data governance on a global scale
3. How the EU can play a leading role in helping start and steer this global trend.

An international group of research funders has been supporting the set-up of the Research Data Alliance (RDA) to enable data exchange on a global scale. The initial phase of RDA has been supported by the European Commission, the US National Science Foundation and National Institute of Standards and Technology, and the Australian Ministry of Research, with research funders from other countries becoming actively involved.²⁷ RDA is being set-up to bring a diversity of stakeholders together and improve interactions between users and technology and service providers.

RDA is a bottom-up community-led initiative to foster global interoperability across geographic and disciplinary boundaries. RDA is open: those who want to participate in RDA and shape the way the global data infrastructure operates are invited to join and take the lead on concrete initiatives. It is focused on the real needs of the research communities and will seek links with industry. It aims at being the place where practitioners stop discussing about the ideal solution and/or the complete set of standards and start implementing practical solutions for data sharing and related issues. Libraries are already active and even in leading roles in several RDA working groups.²⁸

Global initiatives such as COAR (Confederation of Open Access Repositories)²⁹ bring together several major regional repository networks from Australia, Canada, China, Europe, Latin America and the United States. COAR's ambitions are to develop sustainable repository networks all over the globe, align these networks and make them fully interoperable, increase the impact of repository content, and provide training and support. Organisations like EIFL³⁰, World Bank³¹, UNESCO³², ECLAC³³ and others actively promote open access to knowledge as a motor for socio-economic development.

research data itself is global

COAR (Confederation of Open Access Repositories)

Future Horizons

The new European research funding framework reflects the challenges for the next years. Similar examples can be found in other research funding programs in different parts of the world. Horizon 2020, the EU Framework Programme for Research and Innovation, was adopted in December 2013. A quote from the regulation starts: »Horizon 2020 should support the achievement and functioning of the European Research Area in which researchers, scientific knowledge and technology circulate freely, by strengthening cooperation both between the Union and the Member States, and among the Member States, [...]«.³⁴

Horizon 2020 is open also to the participation of non-European countries. The funding scheme includes support for international partnerships e.g. in the domain of scientific information, data and computing-intensive science (areas relevant for COAR, RDA, etc.). Horizon 2020 covers the period of 2014–2020 with a budget of approximately 80 billion Euros. Its macro structure is based on three interrelated pillars: »Horizon 2020 pursues three priorities, namely generating excellent science (›Excellent science‹), creating industrial leadership (›Industrial leadership‹) and tackling societal challenges (›Societal challenges‹). Those priorities should be implemented by a specific programme consisting of three Parts on indirect actions and one Part on the direct actions of the Joint Research Centre (JRC)«.³⁵

Research infrastructures (RI) priority is part of the »Excellent science« pillar and includes e-infrastructures, a.k.a. Information and Communication Technologies Infrastructures offering services for high-speed connectivity, high-performance computing and research data management. It aims at developing a strong European research capacity in terms of instruments, installations and equipment to cope with the most demanding requirements for pushing forward the frontiers of scientific knowledge. The actions on e-infrastructures, as recently published in the Horizon 2020 Work Programme 2014–15, cover data-intensive science and engineering, high-performance computational infrastructure, research and education networks, virtual research environments, and e-science software environments. These actions provide opportunities for partnerships of scholarly communication and data management experts from libraries and scientific communities with e-infrastructure service providers capable of exploring the technologies and know-how for data management supported by high bandwidth communication, high-performance computing, open scientific software, and virtual research environments.

CONCLUSIONS

Libraries' support to research is evolving. The key competency of information provision stays – albeit in an increasingly digital form. But the more tacit role of the library as a service organization that can provide sustainable support for knowledge resources is pushed to the foreground. Text-based resources are complemented by research data. Involvement in digital research methods and operation of software resources becomes a must. Libraries build virtual teams with research offices and computing centers both on a local and a global level and become an integral part of a global e-infrastructure for research.

ACKNOWLEDGEMENTS

The authors would like to thank their colleagues for rich and fruitful discussions.

¹ See COM(2012) 0401 final of 17.7.2012. Available at: <http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1398254671867&uri=CELEX:52012DC0401>

² See http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

³ Kenning Arlitsch, Patrick S. O'Brien (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar, *Library Hi Tech*, Vol. 30, Iss. 1, pp. 60–81. Available at: <http://dx.doi.org/10.1108/07378831211213210>

⁴ <https://www.openaire.eu>

⁵ www.arl.org/news/arl-news/2773-shared-access-research-eco-system-proposed-by-aau-aplu-arl

⁶ <https://www.coar-repositories.org>

⁷ Birgit Schmidt, Iryna Kuchma (2012). Implementing Open Access Mandates in Europe. Available at: www.univierlag.uni-goettingen.de/content/list.php?notback=1&details=isbn-978-3-86395-095-8

⁸ See e.g. the LIBER Quarterly special issue Vol 23, No 4 (2014), <http://liber.library.uu.nl/index.php/lq/issue/view/528>

⁹ Leonardo Candela, Donatella Castelli, Pasquale Pagano (2013). Virtual Research Environments: an Overview and a Research Agenda. Available at: <http://dx.doi.org/10.2481/dsj.GRDI-013>

¹⁰ See Carlos Morais Pires, Jean-Claude Guédon, Alan Blatecky. Scientific Data Infrastructures: Transforming Science, Education, and Society; *Zeitschrift für Bibliothekswesen und Bibliographie* 60 (2013).

¹¹ Susan Reilly, Wouter Schallier, Sabine Schrimpf, Eefke Smit & Max Wilkinson (2011). Report on integration of data and publications, Opportunities for Data Exchange project (ODE). Available at: www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf

¹² See Carole Goble, David de Roure. The Impact of Workflow Tools on Data-centric Research (in Tony Hey et al (ed.). *The Fourth Paradigm: Data-Intensive Scientific Discovery*). Available also at http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part3_goble_deroure.pdf

¹³ See *Advancing Technologies and Federating Communities, A Study on Authentication and Authorisation Platforms for Scientific Resources in Europe: Final Report*. Available at <http://bookshop.europea.eu/en/advancing-technologies-and-federating-communities-pb-KKo413180/>

¹⁴ See Ellen Collins, Michael Jubb. Information Handling in Collaborative Research. *LIBER Quarterly*, [S. l.], v. 22, n. 4, p. 331–344, feb. 2013. ISSN 2213-056X. Available at: <http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-114291/8709>. Date accessed: 24 Apr. 2014.

¹⁵ www.dataone.org/for-librarians

¹⁶ <https://www.force11.org/>

¹⁷ www.orcid.org/

¹⁸ www.crossref.org/fundref/

¹⁹ See Christopher J. Shaffer (2013). The Role of the Library in the Research Enterprise, in: *Journal of eScience Librarianship* 2(1): Article 4. <http://dx.doi.org/10.7191/jeslib.2013.1043>

²⁰ See Herbert van de Sompel, Carl Lagoze. All Aboard: Toward a Machine-Friendly Scholarly Communication System, in: Tony Hey et al (ed.). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Available at: http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part4_sompel_lagoze.pdf

²¹ See http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

²² See Lucie Guibault, Andreas Wiebe (Eds.) (2013). Safe to be Open: Study on the Protection of Research Data and Recommendation for Access and Usage. Available at: www.univerlag.uni-goettingen.de/content/list.php?details=isbn-978-3-86395-147-4

²³ www.arxiv.org

²⁴ www.ncbi.nlm.nih.gov/pubmed

²⁵ <https://www.datacite.org>

²⁶ See F. Friend, H. Van de Sompel, J.-C. Guédon. Beyond Sharing and Re-using: Toward Global Data Networking. Accessed at: <https://eu.ropc.rd-alliance.org/Repository/document/Publications%20and%20Reports/Toward-Global-Data-Networking.pdf>

²⁷ See <https://www.rd-alliance.org/>; Fran Berman. Building Global Infrastructure for Data Sharing and Exchange Through the Research Data Alliance, www.dlib.org/dlib/january14/01guest_editorial.html, doi:10.1045/january2014-berman

²⁸ Beth Plale, »Synthesis of Working Group and Interest Group Activity One Year into the Research Data Alliance«, www.dlib.org/dlib/january14/plale/01plale.html, doi:10.1045/january2014-plale

²⁹ www.coar-repositories.org/

³⁰ www.eifl.net/

³¹ <https://openknowledge.worldbank.org/>

³² <http://en.unesco.org/themes/building-knowledge-societies>

³³ <http://repositorio.cepal.org>

³⁴ REGULATION (EU) No 1290/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 December 2013 laying down the rules for participation and dissemination in »Horizon 2020 – The Framework Programme for Research and Innovation (2014–2020)« and Repealing Regulation (EC) No 1906/2006, Official Journal of the European Union L 347/81, 20.12.2013.

³⁵ Proposal for a COUNCIL DECISION establishing the Specific Programme Implementing Horizon 2020 – The Framework Programme for Research and Innovation (2014–2020) /* COM/2011/0811 final - 2011/0402 (CNS).

THE AUTHORS

Dr. Wolfram Horstmann, Associate Director, Bodleian Libraries, University of Oxford, Broad Street, Oxford OX1 3BG, UK

(Now University Librarian, Göttingen University, Tel.: 0551-39-5210,

E-Mail: horstmann@sub.uni-goettingen.de)

Wouter Schallier, Chief, Hernán Santa Cruz Library, United Nations, Economic Commission for Latin America and the Caribbean (ECLAC), Av. Dag Hammarskjöld 3477, Vitacura, Santiago, Chile,

E-Mail: wouter.schallier@eclac.org

[The author has co-written this article on his personal behalf. Views expressed in this article are not necessarily shared by his organisation]

Jarkko Siren, Technical and Scientific Project Officer, European Commission DG CONNECT, Avenue de Beaulieu 33, 1160 Brussels, Belgium,

E-Mail: Jarkko.SIREN@ec.europa.eu

[The views of the author are his and do not commit the European Commission]

Dr. Carlos Morais Pires, Programme Officer, Coordinator for Scientific Data e-Infrastructures, European Commission, Avenue de Beaulieu 33, 1160 Brussels, Belgium,

E-Mail: carlos.morais-pires@ec.europa.eu

[The views of the author are his and do not commit the European Commission]