‘It’s not about the catalogue, it’s about the data’
Catalogue 2.0: The future of the library catalogue

(Sally Chambers – Ghent Centre for Digital Humanities, DARIAH Belgium)

Does the library catalogue have a future? This was often the first question that people asked me when I was editing *Catalogue 2.0: the future of the library catalogue*. Now, almost 5 years since its publication, the question is; did the predictions in Catalogue 2.0 come true? The quotation in the title of this paper is from Emmanuelle Bermès of the French National Library. She proposed that “it may be time for libraries to start moving beyond the deeply buried data silos that are today’s library catalogues towards freeing bibliographic data from the confinements of the catalogue and making it open, available and reusable as part of the global ‘Web of Data’”. For Bermès, “the real added value of library Linked Data is in its (re)use beyond the library domain” (Catalogue 2.0, xix). The aim of this paper is to explore what has become of the library catalogue since Catalogue 2.0 was published in 2013. Did libraries ‘free their metadata[1]’ and make it openly available on the web for (re)use *beyond the library world*? What is the current state-of-the-art when it comes to bibliographic data? Is metadata ‘enough’ for digital humanities researchers who want to analyse full-text collections using digital tools and methods? What will become of library catalogue in the years to come? These are just some of the issues that this paper seeks to explore.

1) Catalogue 2.0: the State of the Art in 2013

Published in 2013, *Catalogue 2.0: The Future of the Library Catalogue*, set out to provide an overview of the current state of the art of the library catalogue and to look to the future, to see what it may become (Catalogue 2.0, xvii). Working together with a team of key professions in the field of library (meta)data, the state of the art of the library catalogue included the then, relatively new “user-centric way of developing library catalogues”. In her chapter, Anne Christensen described a number of ways of “involving users in an iterative, agile, user-centred development process” with a view to making the library catalogue “into a service that users like and want to use again” (Catalogue 2.0, xviii). Starting from this user-centred model, the book explored technological issues, such as how the application of search engine technologies to library catalogues could improve the search experience for library users (Kinster in Catalogue 2.0, pp.17-36) and the range of “products and services that focus entirely on providing an improved experience in the way that libraries provide access to their collections and services” (Breeding in Catalogue 2.0, p37). A further chapter of the book focused on the *mobile library catalogue* as “a view of a library’s collection, with corresponding services, targeted at customers using mobile devices.” (Koster and Heesakkers in Catalogue 2.0, p65).

The second half the book explored the (meta)data and ways in which libraries could “make bibliographic data work harder”, such “FRBRising your catalogue” with the Functional Requirements for Bibliographic Records (FRBR)[2] as “a user-centred method of modelling the bibliographic universe” (Calleweart in Catalogue 2.0, pp. 93-115). Emmanuelle Bermès from the French National Library, went on to boldly state, that “it’s not about the catalogue anymore, it’s about the data” (Bermès in Catalogue 2.0, p117) introducing the idea that “it may be time for libraries to start moving beyond the deeply buried data silos that are today’s library catalogues towards freeing bibliographic data from the confinements of the catalogue and making it open, available and reusable as part of the global ‘Web of Data’”. For Bermès, “the real added value of library Linked Data is in its (re)use beyond the library domain” (Catalogue 2.0, xix).

In the final section of the book, Karen Calhoun, introduced the idea that “the rise of digital scholarship has a profound impact on the way that libraries deliver services for their users”. Calhoun called for “a fundamental rethinking of the research library service framework” calling for libraries to

[1] Seth Van Hooland and Ruben Verborgh’s ‘Free your Metadata’ initiative provides a practical guide to Linked Data for Libraries, Archives and Museums, see: http://freeyourmetadata.org

[2] http://archive.ifla.org/VII/s13/frbr/frbr1.htm

"to consider collectively new approaches that could strengthen their roles as essential contributors to emergent, network-level scholarly research infrastructures." (Calhoun in Catalogue 2.0, 143)." Is, Calhoun asked, "'catalogue 2.0' a catalogue at all?" (Catalogue 2.0, xix). In the concluding chapter, Dempsey, looked at how the centre of user attention has moved. Previously "users would build their workflows around the library", but Dempsey argues that this is no longer the case, as "users are accustomed to the web and multiple ways of digital delivery, will the library catalogue, describing only part of the 'global collection', remain as an identifiable library service?" (Catalogue 2.0, xx)

So, did libraries "free their metadata" and make it openly available on the web for (re)use *beyond the library world*? What is the current state-of-the-art when it comes to bibliographic data? Is metadata 'enough' for digital humanities researchers who want to analyse full-text collections using digital tools and methods? What will become of library catalogue in the years to come? These are just some of the issues that this paper seeks to explore.

2) Bibliographic Data 2.0

The introduction to Catalogue 2.0 stated that 'the year 2013 will be an important one for libraries. In the United States, implementation of RDA, the new 'cataloguing code' is scheduled for implementation from 31 March 2013. The replacement for MARC 21, for encoding bibliographic data, is likely to be launched and maybe even implemented in 2013.' (Catalogue 2.0, xvii). So, since 2013, how far has the 'bibliographic universe' developed?

Just a few months before the publication of Catalogue 2.0, the Library of Congress announced 31 March 2013 as 'RDA Implementation Day One'[3]. *RDA, Resource Description and Access*[4], intended as the cataloguing standard for the 21st century was "designed for the digital world and an expanding universe of metadata users[5]". Since then, the Library of Congress together with the British Library[6] can be seen the front-runners in RDA implementation, due to its roots in the Anglo-American world. For example, the British Library also implemented RDA from 1 April 2013, as well as coordinating the implementation of RDA with national and international partners. This implementation has since been also rolled out to the UK's Legal Deposit Libraries. In 2012, the British Library's bibliographic systems were reconfigured to support RDA, with the British Library distributing RDA records from 1 June 2012.

Regarding RDA implementation in Europe, the European RDA Interest Group (EURIG)[7], as outlined in the EURIG Cooperation Agreement[8], was established to 'promote the common professional interests of all users, and potential users, of 'RDA: Resource Description and Access', in Europe". The minutes of the EURIG Members meetings are a good source of information regarding the current status of RDA implementation in Europe. For example, at the EURIG Members meeting 2016, RDA implementation had already taken place or was underway in Czech Republic, Finland, Iceland, Italy, Latvia, The Netherlands, Spain (Catalonia only) and the United Kingdom together with collaborative efforts in the German-speaking countries of Austria, Germany and German-speaking Switzerland. Additionally in Denmark, Estonia, Lithuania, Norway, Poland, Slovakia and Slovenia, plus France and French-speaking Switzerland, discussions about possible implementations of RDA were reported[9]. Since 2013, the implementation of RDA has been progressing step-by-step, but as yet, is not widespread in Europe.

[3] http://www.loc.gov/catdir/cpso/news_rda_implementation_date.html

[4] https://www.loc.gov/aba/rda/

[5] http://www.rdatoolkit.org

[6] http://www.bl.uk/bibliographic/catstandards.html

[7] http://www.slainte.org.uk/eurig/index.htm

[8] http://www.slainte.org.uk/eurig/docs/EURIG_cooperation_agreement_2011.pdf

[9] http://www.slainte.org.uk/eurig/docs/EURIG2016/2016_EURIG_Minutes_rev.pdf

Alongside RDA, a related initiative, which was just starting to take shape in 2013, was the Bibliographic Framework Transition Initiative, or BIBFRAME[10]. Initiated by the Library of Congress, BIBFRAME "is an initiative to evolve bibliographic description standards to a linked data model, in order to make bibliographic information more useful both within and outside the library community". As described in Catalogue 2.0, BIBFRAME was intended to replace MARC, Machine Readable Cataloguing, as a standardised way of encoding bibliographic data. Articles describing the challenges of dealing with bibliographic data encoded in MARC have been widely published, following on from Roy Tennant's famous article in the *Library Journal* from 2002, 'MARC must Die'[11]. Since then, similar authors such as Lukas Koster, one of the contributors to Catalogue 2.0, proclaimed 'Who needs MARC?',[12] as well the workshop I organised at the European Library Automation Group 2011 conference, entitled 'MARC must die?'[13]. Since 2013, the BIBFRAME model has been evolving, including active involvement of the community, for example through the BIBFRAME email discussion list[14] and through pilot initiatives[15]. As a result of these consultations, BIBFRAME 2.0 was launched in April 2016[16]. The BIBFRAME website contains a wealth of information about the data model and related vocabularies; various tools, such as the BIBFRAME Editor[17] as well as details of organisations who have or plan to implement BIBFRAME.[18] Furthermore, there are some demonstration datasets and a series of webcasts providing updates on the initiative[19]. However, as the BIBFRAME Frequently Asked Questions clearly state, "BIBFRAME is far from an environment that you could move to yet. The model and its components are still in discussion and development - a work in progress. When it is more mature, vendors and suppliers will need time to adjust services to accommodate it. And then we can expect a mixed environment for some time."[20] Regarding the relationship between RDA and BIBFRAM, interoperability is intended, however, the aim behind BIBFRAME is that it is "independent of any particular set of cataloging rules."

Alongside such large-scale and long-term bibliographic initiatives such as RDA and BIBFRAME, the opening up of (National) Library Metadata has moved on in leaps and bounds. A few notable examples include; the British Library's Free Data Services[21], the Bibliographic Services Data Service of the German National Library[22] and the data.bnf.fr service of the National Library of France[23].
It is not only National Libraries that have been opening up their bibliographic data. Other similar initiatives include Ghent University Library's Open Data[24] and in the United States, a series of *Linked Data for Libraries initiatives* funded by the Andrew W. Mellon Foundation[25] and involving the University Libraries of Columbia, Cornell, Harvard, Princeton and Stanford University as well as the

[10] https://www.loc.gov/bibframe/

[11] http://lj.libraryjournal.com/2002/10/ljarchives/marc-must-die/

[12] http://commonplace.net/2009/05/who-needs-marc/

[13] http://www.slideshare.net/schambers3/marc-must-die

[14] http://listserv.loc.gov/listarch/bibframe.html

[15] https://www.loc.gov/bibframe/docs/pdf/bibframe-pilot-phase1-analysis.pdf

[16] https://www.loc.gov/bibframe/docs/bibframe2-model.html

[17] https://github.com/lcnetdev/bfe/

[18] https://www.loc.gov/bibframe/implementation/register.html

[19] https://www.loc.gov/bibframe/media/

[20] https://www.loc.gov/bibframe/faqs/

[21] http://www.bl.uk/bibliographic/datafree.html

[22] http://www.dnb.de/EN/Service/DigitaleDienste/Datendienst/datendienst_node.html

[23] http://data.bnf.fr/about

[24] http://lib.ugent.be/en/info/open

[25] https://www.ld4l.org

Library of Congress. Most recently, the *Linked Data for Production (LD4P)*, which aims to "begin the transition of technical services production workflows to ones based in Linked Open Data (LOD)" and the *Linked Data for Libraries (LD4L) Labs*, which aims in "helping libraries use linked data to improve the exchange and understanding of information about scholarly resources".

In order to stay in touch with the Linked Open Data (LOD) initiatives in libraries and related organisations, particularly in Europe, the annual Semantic Web in Libraries (SWIB) conference[26] is a must. Starting out as a German language conference in 2009 for 'innovative librarians', it has now grown to a key international library conference, which is held in November each year. At the 2016 conference, Corine Deliot and her colleagues from the British Library's presentation entitled *Who is using our linked data?*[27], may be of particular interest. Their paper reports on the results of a project to examine the usage of the British National Bibliography (BNB) Linked Data Platform[28] through the development of the 'Linked Open Data Analytics platform', which could potentially be of interest to the wider Linked Open Data community.

3) Beyond bibliographic data: from Library Data to Research Data

Emmanuele Bermès's second plea was related to the (re)use of library data *beyond* the library domain. Already by the publication of Catalogue 2.0, the Europeana Data Model (EDM), an "interoperability framework for cultural heritage resources aggregated by Europeana" had been developed (Catalogue 2.0, p128). Since then Europeana has focused on opening up the Europeana data set for (re)use. The Europeana Data Collections[29] are made available by the Europeana Applicaton Programming Interfaces (APIs.)[30] The Europeana Licensing Framework[31], with the aim of standardising and harmonising rights related information and practices, together with the API Terms of Use[32] provides guidelines about how the APIs and related data can be used.

To facilitate the inclusion of library data in Europeana, the Europeana Data Model for Libraries[33] was developed. The aim of EDM for libraries was "to create of a robust aggregation model to make digital content from research and national libraries across Europe available on both Europeana and The European Library portal." The development of the model drew on the expertise of library-domain metadata experts and recommended best practices for aligning library metadata to EDM (Catalogue 2.0, p129). Archives Portal Europe[34] is a similar initiative for archival institutions. Unfortunately, December 2016 saw the closure of The European Library (TEL)[35]. Originally launched by the Conference of European National Librarians (CENL) in 2004 as the union catalogue of European national libraries, it later became a web portal and open data hub for national library data in Europe. Following a consultation among CENL members, it was decided to cease The European Library's services from 31.12.2016[36].

[26] http://swib.org

[27] http://swib.org/swib16/programme.html#abs23

[28] http://bnb.data.bl.uk

[29] http://labs.europeana.eu/data

[30] http://labs.europeana.eu/api

[31] http://pro.europeana.eu/get-involved/europeana-ipr/the-licensing-framework

[32] http://www.europeana.eu/portal/en/rights/api.html

[33] http://pro.europeana.eu/page/europeana-libraries-edm

[34] https://www.archivesportaleurope.net/home

[35] http://www.theeuropeanlibrary.org

[36] See: http://www.cenl.org/wp-content/uploads/20161213-TEL-closure-press-release-final.pdf and http://www.theeuropeanlibrary.org/tel4/newsitem/10000

A further key interoperability framework that has gained traction since the Catalogue 2.0 was published, is the *International Image Interoperability Framework (IIIF)*[37]**.** This collaboratively produced interoperability technology and community framework for image delivery, has a growing community of libraries and cultural heritage intuitions, including several national libraries and leading research libraries.[38] Additionally, Europeana has also aligned with IIIF[39]. Focussing on "image-based resources", the scope of IIIF is wider than one might anticipate, as "digital images are a container for much of the information content in the Web-based delivery of images, books, newspapers, manuscripts, maps, scrolls, single sheet collections, and archival materials"[40]. Ghent University Library has recently implemented IIIF for their image collection[41]. Additionally in Belgium, the PACKED.BE Centre of Expertise in Digital Heritage has been working on the implementation of a Data Hub for museums[42] and the BALaT: Belgian Art Links and Tools[43] from the Royal Institute for Cultural Heritage are examples of data collections from cultural heritage institutions.

Moving beyond the cultural heritage sector, the interest in research data and the role of libraries in *Research Data Management* has increased significantly since the publication of Catalogue 2.0. Institutions such as the Digital Curation Centre in the UK[44], the European Commission[45] and initiatives such as the Research Data Alliance (RDA)[46] and DataCite[47] have taken substantial steps in this area. For example, as part of the European Commissions' Horizon 2020 Research Programme, there is strong advocacy for *Findable, Accessible, Interoperable and Reusable (FAIR)* research data. In the words of the European Commission, "good research data management is not a goal in itself, but rather the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse".

LIBER, the association of European Research Libraries, also urges its members to take on a "leadership role" in the area of research data management[48]. The job of data librarian is becoming increasingly used within the professional literature[49] along with the recognition of *data science*[50]. This area of work presents a huge opportunity for "data professionals" within the library sector. In the coming years, "data libraries" such as the Bodleian Data Library'[51], will become increasingly commonplace. Already, digital research infrastructures such as DARIAH, the Digital Research

[37] http://iiif.io

[38] http://iiif.io/community/#participating-institutions

[39] http://pro.europeana.eu/blogpost/europeana-aligns-with-the-international-image-interoperability-framework-iiif

[40] http://iiif.io/about/

[41] For example, see: http://lib.ugent.be/viewer/archive.ugent.be%3A8ED9BD60-2689-11E6-BB79-D668D43445F2#?c=0&m=0&s=0&cv=0

[42] http://www.packed.be/en/projects/readmore/datahub_voor_musea/

[43] http://balat.kikirpa.be

[44] http://www.dcc.ac.uk/resources/data-management-plans

[45] http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

[46] https://www.rd-alliance.org/about-rda

[47] https://www.datacite.org

[48] http://libereurope.eu/wp-content/uploads/2015/11/LIBER-Libraries-and-Research-Data-factsheet1.pdf

[49] For example, see: Rice, R. and Southall, J. (2016), *The Data Librarian's Handbook***,** London, Facet Publishing: http://www.facetpublishing.co.uk/title.php?id=300471#.WJnm5BiZPVo

[50] For example, the recent launch of the Data Science Journal: http://datasciencehub.net

[51] https://www.bodleian.ox.ac.uk/data

Infrastructure for the Arts and Humanities, are exploring the 'fluidity' of data within their infrastructures[52]. Within Belgium, a feasibility study for a Social Sciences and Humanities Data Archive (SODAH)[53] has recently been funded by the Belgian Science Policy office and is another very interesting development in the area of data archiving and reuse.

4) Digital Collections to Digital Scholarship

While many researchers acknowledge the importance of high quality metadata to enable the discoverability of the resources they need for their research, metadata is often seen as a means to an end, rather than an end in itself. If we define data in the widest possible sense as a collection of bits and bytes, then the *data* referred to in the title of this paper can include not only *metadata* or "data about data", but also the data itself. As cultural heritage institutions increasingly digitise their collections, they are in effect, converting their collections into *data*. Particularly in the area of digital humanities, the analysis of this full-text data is becoming increasingly important. For example, the National Library of the Netherlands recently organised a workshop entitled "Historical Newspapers as Big Data".[54] The focus of this workshop was to bring together researchers from a range of disciplines who were interested in using the digitised newspapers and other digital collections made available via the Delpher platform[55] for (digital) humanities research. During the workshop, the National Library's new service of providing access to the 'Data in Delpher' was announced[56]. This data includes 111GB of text from the Delpher Open Newspaper archive[57] from 1618-1876. Shortly, texts from the Book and Periodical collections (1850-1876) as well as radio bulletins (1930 – 1984) from the ANP (Algemeen Nederlands Persbureau) will also be available shortly. Additionally, you can contact the "data services team" directly, for specific requests. The provision of this kind of access to the *data* of the digital library collections will be increasingly important.

Within the framework of DARIAH-BE, Ghent Centre for Digital Humanities (GhentCDH) is leading the development of a Digital Text Analysis platform. This work is funded by the Research Foundation Flanders (FWO) as part of the DARIAH Virtual Research Environment Service Infrastructure (VRE-SI). The aim of the platform is to enable text analysis on digitised textual collections (e.g. digitised newspapers, digitised books or even web-archives) for use by a wide-range of (digital) humanities projects. The digital textual analysis platform will enable researchers to browse and search the full-text of digitised collections. Once the relevant research sub-corpus has been identified, data export tools will enable researchers to export sub-corpus, in standard open formats (such as XML, JSON, .csv, .txt, etc) for analysis with existing digital text analysis tools such as MALLET, (http://mallet.cs.umass.edu/topics.php) for *topic modelling*, VOYANT (http://voyant-tools.org) for *data visualisation* or AntConC (http://www.laurenceanthony.net/software/antconc/) for *concordance* and *textual analysis*. The possibility of including part of the Google Books collection at Ghent University Library in the digital text analysis platform is currently being investigated.[58] Furthermore, legislative changes are needed to enable text and data mining on libraries digital collections. The European Association of Research Libraries, LIBER is actively advocating for more flexible copyright system that will allow text and data mining to be used to its full potential[59].

[52] Romary, L., Mertens, M and Baillot, A (2016) *Data fluidity in DARIAH – pushing the agenda* forward in *BIBLIOTHEK Forschung und Praxis*, De Gruyter, 2016, 39 (3), pp.350-357: https://hal.inria.fr/hal-01285917

[53] https://cessda.net/National-Data-Services/CESSDA-Members/Belgium

[54] https://www.kb.nl/nieuws/2016/historische-kranten-als-big-data-ii-concepten-op-drift

[55] http://www.delpher.nl

[56] http://www.delpher.nl/nl/platform/pages/helpitems?nid=482

[57] Delpher Open Newspaper Archive (1.0). Creative Commons Attribution 4.0, The Hague, 2017 and http://www.delpher.nl/nl/platform/pages/helpitems?nid=513&scrollitem=true

[58] http://www.ugent.be/en/facilities/library/google-project.htm

[59] http://libereurope.eu/text-data-mining/

In the coming years, advances in Handwritten Text Recognition (HTR) technology will improve the ability of software to automatically recognise handwritten texts. For archives and libraries, with many handwritten documents in their collections, these advances will significantly increase digital access to these documents. Initially funded under the European-funded project Transcriptorium[60] and more recently in a follow-up project, READ (Recognition and Enrichment of Archival Documents) project[61], the aim is "to revolutionize access to archival documents with the support of cutting-edge technology such as Handwritten Text Recognition (HTR) and Keyword Spotting (KWS)." A key element of these projects have been the iterative development of the Transkribus platform[62] into a fully functioning service platform for the ''for the automated recognition, transcription and searching of historical documents." While at present using the Transkribus platform to automatically recognise handwriting still requires a significant amount of manual transcription, as use of the platform increases and the underlying algorithms are improved, it is likely that Handwritten Text Recognition will develop into a core technology, in the same way, that Optical Character Recognition (OCR) is widely used in libraries and archives today.

While until now we have talked about the digitisation of analogue resources, in the coming years, there will be an increased amount of 'born digital' material in the collections of libraries and archives. A particularly important born-digital resource are the increasing number of web-archives[63] that are in the process of being developed throughout Europe, for example, in the UK Web Archives[64], the web archives in France at the French National Library[65] and French National Audiovisual Institute (INA)[66] and the Danish Netarchive.[67] A particular challenge is the provision of research access and use of these webarchives.[68] However, initiatives such as RESAW, a Research Infrastructure for the Study of Archived Web Materials[69], are addressing. Although the .be domain was introduced in June 1988, the Belgian web is currently not systematically archived. As of February 2017, 1.562.460 domains are registered by DNS Belgium[70]. Without a Belgian web archive, the content of these websites will not be preserved for future generations and a significant portion of Belgian history will be lost forever. In December 2016 a pilot web archiving project was funded by the Belgian Science Policy Office, BELSPO, called PROMISE (PReserving Online Multiple Information: towards a Belgian StratEgy) that wants to (i) identify current best practices in web-archiving and apply them to the Belgian context, (ii) pilot Belgian web-archiving, (iii) pilot access (and use) of the pilot Belgian web archive for scientific research, and (iv) make recommendations for a sustainable web-archiving service for Belgium[71]. Web-archives, like other born-digital resources, will be core digital collections enabling digital scholarship.

[60] http://transcriptorium.eu

[61] https://read.transkribus.eu

[62] https://transkribus.eu/Transkribus/

[63] https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives

[64] For an overview of web-archives in the UK see: Winters, J. (2016) *Negotiating the archives of the UK webspace* at Workshop on National Webs, Aarhus, 8-9 December 2016: http://www.netlab.dk/wp-content/uploads/2016/12/12-Jane-Winters-National-Webs-Workshop-December-2016.pdf

[65] http://www.bnf.fr/en/collections_and_services/book_press_media/a.internet_archives.html

[66] http://www.inatheque.fr/fonds-audiovisuels/sites-web-media.html

[67] http://netarkivet.dk/in-english/

[68] Truman, Gail. 2016. Web Archiving Environmental Scan. Harvard Library Report: https://dash.harvard.edu/handle/1/25658314

[69] http://resaw.eu

[70] https://www.dnsbelgium.be

[71] Chambers, S et al. (2016) *Towards a national web in a federated country: a Belgian case study* at Workshop on National Webs, Aarhus, 8-9 December 2016: http://www.netlab.dk/wp-content/uploads/2016/12/09-Sally-Chambers-20161209_Towards_National_Web_Belgium.pdf

5) Libraries as partners in digital humanities research

The role of libraries in digital humanities has been a topic for discussion for several years[72]. However, step-by-step, with the increasing number of initiatives in this field, such as the Alliance of Digital Humanities Organizations (ADHO)'s *Libraries and the Digital Humanities Special Interest Group*[73] and the *DH+Lib community*[74], this role is becoming more clearly defined. Within Europe, flagship examples of digital scholarship activities in libraries are the *Digital Research team* at the British Library[75] as well as the Digital Humanities team at the National Library of the Netherlands[76].

At the LIBER (European Association of Research Libraries) Annual conference in 2015, DARIAH and LIBER organised a joint workshop on 'From Digital Collections to Digital Scholarship: new takes on research support'. This workshop explored questions such as: "How can librarians gain a better understanding of what humanities researchers need when interacting with data, creating data, annotating and mining large and complex digital collections?", "How can libraries build on their existing strengths to offer valuable expertise for digital humanities projects?" and "What other partnerships are needed, e.g. university/faculty IT services?". Building on the results of this workshop, LIBER and DARIAH decided establish two working groups with the goal of facilitating existing activities and encouraging interested libraries to engage with this growing group of humanities scholars.

The *LIBER Digital Humanities Working Group*[77] was formally launched in February 2017 and is co-chaired by Andreas Degkwitz, Chief Librarian of Humboldt University, and Lotte Wilms, Digital Scholarship Advisor at the National Library of the Netherlands. The LIBER Working Group will focus on creating a knowledge network of libraries around Digital Humanities (DH) within Europe. The goal of this network is to facilitate knowledge sharing, improving services for the academic community and thereby strengthening the relationship of European research libraries with digital scholars. The working group will connect people and organisations and act as a platform for exchanging practical examples and experiences, as well as a discussion group for strategic themes surrounding Digital Humanities.

The *DARIAH Working Group* is in the process of being formally established, following its first meeting at the DARIAH-EU Annual Event in Ghent in October 2016. The aim of the DARIAH-EU Working Group is to explore opportunities to strengthen collaboration between researchers and libraries in the area of digital humanities by exploring how the expertise of both can be maximally deployed in digital humanities projects. By combining each other's strengths, the working group will look into successful collaborations and identify case studies that can be used as a reference for future projects. A balanced membership between digital humanities researchers from a range of humanities disciplines / methodological approaches and library staff is critical to the success of the Working Group.

While the two working groups have separate focus areas, the groups will work closely together and where possible will organise joint activities. For example, the working groups will jointly organise a session at the International Federation of Library Associations and Institutions (IFLA) Satellite

[72] For example, see Miriam Posner's Digital Humanities and the Library bibliography: http://miriamposner.com/blog/digital-humanities-and-the-library/

[73]https://docs.google.com/forms/d/e/1FAIpQLSfswiaEnmS_mBTfL3Bc8fJsY5zxhY7xw0auYMCGY_2R0MT06w/viewform

[74] http://acrl.ala.org/dh/

[75] https://www.bl.uk/subjects/digital-scholarship

[76] https://www.kb.nl/en/organisation/research-expertise/digital-humanities

[77] http://libereurope.eu/blog/2017/02/03/liber-launches-digital-humanities-working-group/

Meeting *Digital Humanities – Opportunities and Risks: Connecting Libraries and Research*[78] which will take place in Berlin in August 2017. The activities of these working groups are intended to address questions and facilitate opportunities for researchers and libraries to work together in the area of digital humanities.

Archives, just as much as libraries have a role to play in digital humanities research[79]. Within in DARIAH, a DARIAH-BE initiated Working Group on the *Sustainable publishing of archival catalogues*[80] has been established. The aim of this Working Group is to "offer a platform for archivists to reflect on their methodologies to ensure sustainable publishing and sharing of their metadata and data". In this way, the challenges faced by archives contributing to digital humanities initiatives can be addressed an overcome.

6) Big Data in the Arts and Humanities?

For digital humanities to truly reach its potential, strong collaboration is needed between (digital) humanities researchers, cultural heritage institutions and computer scientists. In summer 2013, the UK's Arts and Humanities Research Council launched a call for Big Data Projects[81] as part of their Digital Transformations programme[82]. The Digital Transformations programme aims "to exploit the potential of digital technologies to transform research in the arts and humanities" and particularly to tackle "crucial issues such as intellectual property, cultural memory and identity, and communication and creativity in a digital age." A total of 21 projects, from a diverse range of subject areas such as "Standards for Networking Ancient Prosopographies: Data and Relations in Greco-roman Names", "Visualising European Crime Fiction: New Digital Tools and Approaches to the Study of Transnational Popular Culture", "A Big Data History of Music" and the "Big UK Domain Data for the Arts and Humanities (BUDDAH)". This initiative helped to demonstrate the potential of big data technologies for arts and humanities research.

Another similar initiative, the 'Research Data Spring[83], organised by the UK's Joint Information Systems Committee (JISC), focussed on supporting 'the creation of innovative partnerships between researchers, librarians, publishers, developers and other stakeholders engaged in the research data lifecycle'. A particularly interesting project funded under this programme was the *Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections*[84]. In this project, humanities scholars worked together with colleagues from the British Library and computer scientists from the High Performance Computer Centre (HPC) at University College London, to analyse how HPC services could be used with 'humanities research data sets'. The humanities dataset chosen was a "60,000 book dataset covers publication from the 17th, 18th, and 19th centuries, or – seen as data – 224GB of

[78] https://dh-libraries.sciencesconf.org

[79] See for example: Petra Links and Reto Speck. "The Missing Voice: Archivists and Infrastructures for Humanities Research." *International Journal of Humanities and Arts Computing*. 7.1-2 (Oct. 2013). http://dx.doi.org/10.3366/ijhac.2013.0085

[80] The DARIAH-EU Working Group, Sustainable Publishing of Archival Catalogues was established following a series of workshops funded under the DARIAH-EU Open Humanities theme, see: Vanden Daelen, V. et al (2016) *"Open History: Sustainable digital publishing of archival catalogues of twentieth-century history archives"*, Dec 2015, Brussels, Belgium. 2016: https://hal.inria.fr/hal-01281442

[81] http://www.ahrc.ac.uk/funding/opportunities/archived-opportunities/bigdataprojectscall/

[82] http://www.ahrc.ac.uk/research/fundedthemesandprogrammes/themes/digitaltransformations/

[83] See: https://www.jisc.ac.uk/rd/projects/research-data-spring and https://researchdata.jiscinvolve.org/wp/category/research-at-risk-2/research-data-spring/

[84] Williams, O., Farquhar, A. (2016). Enabling Complex Analysis of Large-Scale Digital Collections: Humanities Research, High Performance Computing, and transforming access to British Library Digital Collections. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 376-379: http://dh2016.adho.org/abstracts/230

compressed ALTO XML that includes both content (captured using an OCR process) and the location of that content on a page"[85]. Working together with humanities researchers, HPC colleagues 'translated' the humanist's research questions into computational queries that could be used on the book data set. For example, one of the research case studies in the project explored whether there was a correlation between existing demographic data on known epidemics of various infectious diseases (e.g. Cholera, Whooping Cough, Consumption, and Measles) and an increase in related vocabulary in books published in the same years as known epidemics. A particularly interesting recommendation from the research was the possible role for librarians to work with researchers to 'translate' their research questions into computational queries. As noted in the article, "training librarians to aid humanities researchers in carrying out defined computational queries via adjustable recipes would improve access to infrastructure, and cut down on the human-resource intensive nature of this approach". It would be very interesting to explore the possibility of undertaking a similar exercise in Belgium.

Another interesting example of the use of high-performance computing in the arts and humanities is the 'Cultural Heritage Cluster' at Aarhus State and University Library[86]. Developed as part of a Danish eInfrastructure Cooperation initiative, with the aim of spreading High Performance Computing (HPC) to new research areas such as the humanities and social sciences, the *DeIC National Cultural Heritage Cluster, State and University Library* "applies state-of-the-art technologies within data science, and for the first time ever facilitates quantitative research projects on the digital Danish cultural heritage." In 2016, a call for pilot projects from the humanities and social sciences was launched in which researchers were offered access and use of the *Cultural Heritage Cluster*, including related training. Initially, three pilot projects have been selected: (a) *Probing a Nation's Web Domain*, a study of the Danish web archive, (b) *Digital Footprints* a research project analysing social media data and (c) a project to *analyse the development in the Danes' language usage on the social media*. Such projects offer much inspiration as to the potential of large scale computing for digital humanities research.

In autumn 2016, at the invitation of the Belgian Science Policy Office, Belspo, DARIAH-BE was invited to organise a workshop on *Fostering closer collaboration: e-Infrastructures and DARIAH-BE.*[87]The aim of the workshop was to a) increase the understanding of Arts and Humanities researchers of what Belgian eInfrastructures (such as High Performance Computing) offer, b) increase the understanding of the e-Infrastructural needs of Arts and Humanities researchers and c) to identify concrete collaboration actions between the two communities in 2017-2018. Already, at the *High Performance Computing Infrastructure* at Ghent University[88] one the researchers affiliated with Ghent Centre for Digital Humanities (GhentCDH) has been using Ghent's supercomputer to carry out Social Network Analysis on European Parliamentary Debates[89]. Increasing "big data" research in the arts and humanities is something that we would like to develop further in the context of DARIAH-BE.

7) Concluding Words: Priorities for 2020

In conclusion, I would like to return to the aim of this colloquium, Inside the User's Mind and the role of the MADDLAIN project in better understanding the digital practices and needed of the users of archive centres and libraries. Drawing on both my experience of working in the area of interoperability of bibliographical (meta)data, digital research infrastructures and more recently in the

[85] http://dh2016.adho.org/abstracts/230#ftn8

[86] https://en.statsbiblioteket.dk/kulturarvscluster

[87] https://docs.google.com/document/d/1euU2AkbLqYZks9eK6NSrZJUI8g_JvwmePDMi2FzdnFo/edit?usp=sharing

[88] http://www.ugent.be/hpc/en

[89] See: Brankovic, Jelena, Julie Birkholz, and Martina Vukasovic. "Is the Europe of Knowledge the Talk of the Town? Higher Education in the European Parliament." *Consortium of Higher Education Research (CHER) Annual Conference*. 2015. Print. https://biblio.ugent.be/publication/8050481 and https://twitter.com/GhentCDH/status/823894646787031040

digital humanities, I would like to propose some priorities for cultural heritage institutions and libraries in particular, to further develop their services in order to fully contribute to digital research in the arts and humanities:

- **Priority one: Data-level access to the digital collections**: “it’s not about the catalogue, it’s about the data” - the provision of *data-level access*[90] to the digitised and born-digital collections of cultural heritage institutions is *crucial* for digital humanities research. Without this ‘data’ it is simply not possible to undertake digital humanities research. The danger is that a parallel ‘digitisation’ track is established, where researchers simply digitisation what they need themselves, with little consideration for preservation, copyright and other related issues. It would be far better to join forces and together work on providing data-level access to digital content in formats that can be readily fed into the digital research tool of choice[91] of the researcher.

- **Priority two: establish Digital Humanities Centres**: it would be wonderful if a true ecosystem of digital humanities centres were to be established across Belgium[92] to facilitate digital research in the arts and humanities. These centres could include ‘labs initiatives’ [93] where researchers are encourage to ‘experiment’ with a libraries digital collections, including collaborations with High Performance Computer centres[94] or testing out handwritten text recognition software[95]. Such Digital Humanities Centres could offer training programmes[96]**,** research fellowships[97] or placements for PhD students[98] with regular blog posts and tweets on the use of the libraries collections for digital humanities research[99].

- **Priority three: Open Bibliographic Data**: while I have stressed the importance of data-level access to digital collection, the bibliographic (and archival) metadata that describes those collections remain essential. It would be wonderful to see such data in Belgium being openly available as Linked Open Data. Initiatives such as the Linked Open Data instance of the British National Bibliography[100] or data.bnf.fr could provide inspiration and guidance in this area. Could for example, http://data.kbr.be be a reality by 2020?

So, indeed, it is not a choice, between the catalogue or the data, it’s simply about both.

[90] http://www.delpher.nl/nl/platform/pages/helpitems?nid=482

[91] See for example, the DIRT (Digital Research Tools) Directory: http://dirtdirectory.org

[92] Chambers, S., Deroo, K., Dozo, B., Gheldof, T. (2016). DARIAH-BE: Towards an ecosystem of Digital Humanities Research Centres in Belgium, Digital Humanities Centres: Experiences and Perspectives, 8-9 December 2016, Digital Humanities Laboratory (University of Warsaw), Warsaw, Poland, http://hdl.handle.net/2268/203796, see also: https://twitter.com/DARIAHbe/status/824364561650380801

[93] As inspiration: http://labs.bl.uk and http://lab.kbresearch.nl

[94] As inspiration: https://en.statsbiblioteket.dk/kulturarvscluster

[95] As inspiration: https://transkribus.eu/Transkribus/

[96] As inspiration: McGregor, N., Ridge, M., Wisdom, S., Alencar-Brayner, A. (2016). The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 623-625: http://dh2016.adho.org/abstracts/178

[97] As inspiration: https://www.kb.nl/en/organisation/kb-fellowship

[98] As inspiration: https://www.bl.uk/news/2016/november/british-library-phd-placements-call-for-applications

[99] As inspiration: http://blog.kbresearch.nl and http://blogs.bl.uk/digital-scholarship/ and https://www.arts.kuleuven.be/home/nieuwsbrief/archief/2016-2017/onderzoek_in_de_kijker/tom-willaert-digital-scholarship

[100] As inspiration: http://bnb.data.bl.uk and http://data.bnf.fr