Introducing the LEMC: How to build an Early Music Research Infrastructure

Marnix van Berchum
Utrecht University, ICON
Muntstraat 2A
NL-3512 EV Utrecht
0031-6-48794541
m.vanberchum@uu.nl

ABSTRACT

This paper outlines a Research Infrastructure for the study of Early Music. The Large Hadron Collider (LHC) of CERN serves as example for the building blocks needed. The paper discusses the elements of the proposed Large Early Music Collider (LEMC), including the requirements of encoded music, the availability of tools for music analysis, and the position of different types of libraries in the infrastructure. Where relevant initiatives in digital infrastructure and tools, in and outside of musicology, are introduced.

Categories and Subject Descriptors

J.5 **[Arts and Humanities]**: Performing arts (e.g., dance, music).

General Terms

Design, Standardization.

Keywords

Research infrastructures, Early Music, music data, digital tools.

1. THE IMAGINATIVE LHC

One of the largest existing research infrastructures in the world is also one of the most imaginative. Its most impressive part stretches 27 kilometres under territories of two European countries, operates at -271.3°C and is host to a 7000-tonne research instrument. It has a name to fit: the Large Hadron Collider (LHC). Its location: the place where the internet was invented, the CERN near Geneva, Switzerland. In this 'circular accelerator' sub-atomic particles are accelerated to nearly the speed of light, to eventually collide inside large 'detectors'. The results of these collisions – approximately 600 million per second, leading to a staggering 15 petabytes of data each year – provide humanity insight into the fundamental structure of the universe.[8]

Comparing scientific disciplines – with their own questions, research data and methodologies – might not always be fruitful, but infrastructures like the LHC built by and for the Particle Physics community do tickle the mind to find analogies in your own field of research. In this paper I attempt to describe an infrastructure for the study of Early Music. I am in no way an expert or even knowledgeable in the field of particle physics, but I would like to use the components of the LHC as building blocks for the *LEMC: the Large Early Music Collider*. What would such an infrastructure for early music look like? Does musicology need distances of 27 kilometres and temperatures of -271.3°C? Can musicology have ten thousand scholars working on one experiment, and co-authoring articles on the results?

2. PARTICLES OF EARLY MUSIC

But let's start with the smallest parts in the infrastructure, the Early Music particles (EM-particles). Considering the research tasks of the Early Music scholar three types of EM-particles are distinguished in the LEMC:

1) images of Early Music sources
2) encoded music files of Early Music
3) secondary literature on Early Music

All three types are indispensable for performing scholarly study on Early Music, and thus all will be present in the infrastructure to 'collide' with each other. These particles exist in a variety of formats and types, so defining the common requirements to which they should all comply is essential for the research infrastructure to work. A first feature is the open character of the file format; we do not want files in the system that are proprietary to a particular software package, but rather those that can be used as needed. The format and standards used should also be interoperable and flexible. These two requirements are best explained with the example of encoded music files. The *Josquin Research Project* for example uses the long-standing Humdrum/**kern format as basic format in their database, while providing several other formats, like MusicXML and MEI, and also rendered scores in PDF.[7] The possibility of converting the **kern scores to other formats show the strength of this format. In the LEMC all formats used will be able to interoperate in such way with each other, avoiding the choice of one format only. The infrastructure will provide the necessary tools for conversion. An encoding schema for Early Music should be flexible in the sense that if we discover a new dialect of mensural notation the schema should be able to harbour this dialect. A last characteristic of the EM-particles flying around in the LEMC is trust. A user of the infrastructure should be assured of the integrity and authenticity of the files she is using. To that end the infrastructure will provide provenance information of all particles.

DLfM '14, September 12 2014, London, United Kingdom

ACM 978-1-4503-3002-2/14/09 $15.00.

http://dx.doi.org/10.1145/2660168.2660180

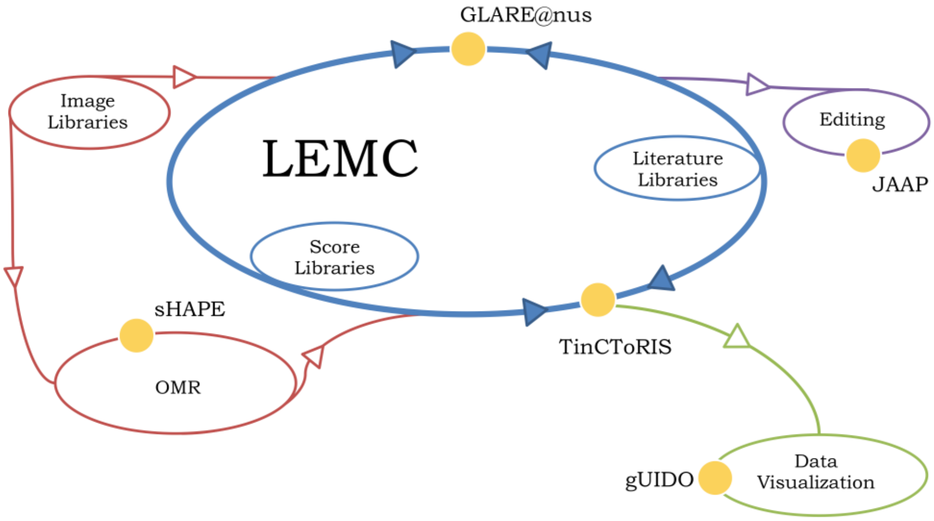

Figure 1. Conceptual structure of the Large Early Music Collider (LEMC). Based on illustration by CERN.[8]

3. THE STRUCTURE OF THE LEMC

Figure 1 shows the conceptual structure of the LEMC, consisting of several circular accelerators with a certain type of content (the open circles), and collision tools (the yellow dots). The main LEMC accelerator is connected to several smaller ones, each with their own role in the infrastructure. The particles can move in and out of them (the direction of the arrows).

3.1 Circular accelerator – music data cloud

The circular circuit in which the particles fly around is, in the case of the LHC, a truly physical circle. In the case of the LEMC this would rather be a virtual circle, but still, it has to be a proper environment where the EM-particles can roam around, can be introduced to or escape from. Above all it should be a place where the particles can be studied. Without going to deep into the discussion on what it really is, one might think of the LEMC as a cloud infrastructure. All partners involved will be able to upload and access their material in the LEMC Cloud. Music Digital Libraries will be the prime partners in providing the content to the LEMC Cloud. The particles these libraries provide can be of one or more of the described types. DIAMM for example will be one of the image libraries, while Europeana provides both images as well as score files.

Like the EM-particles, the cloud structure that hosts them, should comply with the same characteristics of openness, interoperability, flexibility, and trust. One can imagine that the compliance rate of each providing library with the individual requirements may vary. The openness of an image library (i.e. the level of accessibility and re-use of the content, apart from the openness of the format used) can be different to that of a score library. This may differ even within the content provided by one library, for example fully open low-resolution images versus high-resolution images with a login mechanism. All metadata, including provenance information, will be openly available. In that sense the LEMC accelerators will be transparent in what they do.[5, 6]

3.2 Collision tools

Like the experiments hosted on CERN's LHC, the LEMC has different 'detectors' connected to the cloud structure of musical data. The main collision instruments are *GLARE@nus* and *TinCToRIS*, which will offer functionality for the analysis of music and metadata respectively. In the case of *GLARE@nus* the Python-based toolkit Music21 with its current functionalities is an appropriate candidate to be (part of) this instrument. Music21 is able to import the Humdrum/**kern, MusicXML, MIDI, and ABC formats, which all comply with the requirements of the LEMC particles as discussed above.[9] *TinCToRIS* on the other hand will work with the metadata present in the LEMC cloud, for example for finding patterns in the dissemination of music. The outputs of both instruments may feed back into the music data cloud, enriching the material already present, or can be used as finished products outside of the LEMC. The instruments, and their underlying software, all comply with the rules of openness, interoperability, flexibility, and trust.

3.3 Pre and post-accelerators

CERN does not use the LHC only for accelerating articles. It is actually a combination of several linear and circular accelerators, making sure the particles reach the collision speeds needed when colliding in the detectors. The LEMC also hosts several accelerators with their own tools outside of the main ring. As mentioned earlier these (i.e. the circles in Figure 1) should be understood as an environment where a particular type of content is available, re-usable and editable before being transferred to another part of the infrastructure.

The *sHAPE* pre-accelerator exemplifies this. It hosts an Optical Music Recognition (OMR) tool, which uses the files provided by the image libraries as input. Its output consists of encoded music files that enter the main ring of the LEMC, for example to be analysed in *GLARE@nus* and *TinCToRIS*. This part of the LEMC already exists: the Aruspix project provides a software environment for the OMR of early typographic music prints of the 16th and 17th centuries.[1]

Similarly *JAAP* and *gUIDO* are post-accelerators, making use of the output of the main ring of the LEMC. *JAAP* provides high quality tools for making scholarly editions of Early Music. *gUIDO* is an environment in which the output of the metadata analysis done by *TinCToRIS* (but it can as well be *GLARE@nus*) is transformed into data visualisations. Like the pre-accelerators, *JAAP* and *gUIDO* can make use of existing tools. The CMME project offers functionality fitted for *JAAP*, and in the *gUIDO* accelerator tools like Gephi (network visualisations) and GLAMMap (geographical visualisations) can be plugged in.[2, 3, 4]

4. CONCLUSION AND REMARKS

The proposed design provides a direction in which the current existing digital tools useful for the study of Early Music, might integrate with the content present in digital music libraries to form a proper research infrastructure for Early Music. Furthermore it shows the potential of how a small and perhaps less technology oriented discipline can learn from the high tech infrastructures that are around.

This 'vision' is in no way complete though. One might think of different type of particles (e.g. audio files), which are not included and need different accelerators and tools. Moreover, Figure 1 is merely a conceptual view of the LEMC. Other, more detailed designs will shed light on the technical interfaces – and associated challenges – to be built between the different components. Figure 2 shows a first step in defining how the LEMC could look like from an architectural perspective, showing the central data cloud with content providers and research tools accessing it via API's.

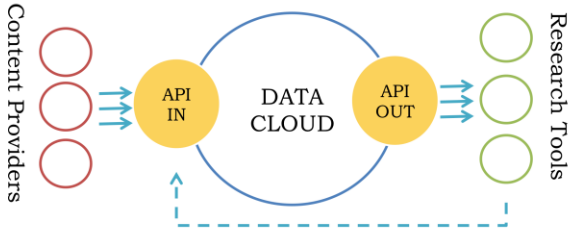

Figure 2. Architectural perspective on the LEMC

The components of the LEMC offer functionalities for many steps in the research process of the Early Music scholar, but the current design stops where publishing starts. The output of the LEMC components will be the input of the publication process. It would be interesting to think about how the LEMC components can be integrated further in this process. What happens if we consider the particles themselves as scholarly output? Would we for example be able to publish the EM-particles straight from the main accelerator circle?

A final issue that is not resolved in this paper is the position of the user of the LEMC. Thinking of the structure of something like the LEMC is perhaps the easy step. Making sure the designated community can effectively use it is something else. Besides 'convincing' and 'educating' scholars to use digital tools, the tools should be developed with a user-centred design approach. The Europeana Cloud project is experimenting with building tools for humanities scholars, on top of the Europeana content. First results were achieved with a group of philosophers studying the history of philosophical concepts.[10] Looking outside one's own discipline might prove to be useful for musicology in this respect too.

5. ACKNOWLEDGMENTS

The idea of a LEMC first came to my mind after a visit to CERN, part of the social programme of the OAI7 Conference in 2011. Over the past couple of years I discussed the notion of a (Early) Music Research Infrastructure with several colleagues. I would like to acknowledge Magchiel Bijsterbosch (SURF), Alastair Dunning (The European Library), Maarten Hoogerwerf (KNAW-DANS), Laurent Pugin (RISM Switzerland), Maurice Vanderfeesten (TU Delft) and Frans Wiering (Utrecht University) for the input they provided on my thoughts on how such a research infrastructure – and research infrastructures in general – would and should look like.

6. REFERENCES

[1] *Aruspix*. http://www.aruspix.net.

[2] *The Computerized Mensural Music Editing Project (CMME)*. http://www.cmme.org.

[3] *Gephi: The Open Graph Viz Platform*. http://www.gephi.org.

[4] *GLAMMap*. http://axiom.vu.nl/GLAMMap.html.

[5] High-Level Group on Scientific Data 2010. *Riding the wave - How Europe can gain from the rising tide of scientific data*. European Union. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

[6] Hogenaar, A., Tjalsma, H., and Priddy, M. 2011. Research in the Humanities and Social Sciences. In Meier zu Verl, C., and Horstmann, W. (Eds) 2011. *Studies on Subject-Specific Requirements for Open Access Infrastructure*. Bielefeld, Universitätsbibliothek. DOI= http://doi.org/10.2390/PUB-2011-1.

[7] *The Josquin Research Project*. http://josquin.ccarh.org.

[8] Lefevre, C. 2009. *CERN-Brochure-2009-003-Eng LHC: the guide (English version)*. CERN Geneva. http://cds.cern.ch/record/1165534?ln=en.

[9] *music21: a toolkit for computer-aided musicology*. http://web.mit.edu/music21.

[10] Van den Berg, H., Parra, G., Jentzsch, A., Drakos, A. and Duval, E. (accepted). Studying the History of Philosophical Ideas: Supporting Research Discovery, Navigation, and Awareness. Paper. *i-KNOW '14*, September 16-19 2014, Graz, Austria. DOI= http://dx.doi.org/10.1145/2637748.2638412.