Oceans of data

Christian-Emil Smith Ore
Unit for Digital Documentation, University of Oslo

The conference theme of CAA2016 is "Exploring oceans of data", hinting at the vast amount of digital data resulting from digitization projects and from all kind of electronic measuring gadgets used to document excavations and surveys. A quick look at this year's conference book of abstracts will tell you that only a minority of the presentations actually address issues connected to curation, organization and use of the oceans of data. The majority of the presentations is, as at all CAA meetings, about innovative and experimental use of computer in archaeology and about the application of existing technology to new scientific projects, that is, about activities producing even more data. This is not unexpected. The system for academic credits gives little or no award for the use and development of research infrastructure.

My first CAA conference was in 1996 in Iasi, Romania. Back then I lead a large digitization and database project for collections and archives build up by scholarly field work since the first part of the 19th century, including the collections of the archeological museums in Norway. The overarching data model was inspired by the event oriented model developed at the Danish National museum in 1988-89 and the data format was based on TEI (Text Encoding Initiative) developed by text philologists from 1987 onwards. As a natural consequence of this interdisciplinary modelling work I eventually got involved in the development of the CIDOC–CRM which is designed to be a conceptual model for data integration.

In 2012 CAA celebrated its 40th anniversary. The CAA2012 had a special session called "personal histories" where key members shared their CAA memories. The session was captured on video, can be viewed online and is highly recommended. Most of the memories are about social events and about the primitive state of computers back then, as it should be. However, there were a few caveats. In two of the 2012 "personal histories" it was stressed that we must not forget about proper archiving and that there is no point in storing your data in the cloud if you cannot read them after a few years.

Paper based data are voluminous and less accessible than digital data but are stable and can after years eventually find its way to collections and archives. Digital data are fragile and will usually not be readable after years in the attic. Without proper actions, the floods of digital data may evaporate and the oceans of data shrink as an Aral Sea.

Archaeology is neither library nor archival science. But a substantial part of archaeological training is how to do sound and accurate documentation of contexts. Methods for construction, curation and reuse of archaeological datasets should be in the central focus as well. Standardized conceptual data models can ease curation and secure long term reusability. Used for these purposes models will not put straitjackets on research.

Under the assumption that we manage to create and preserve the oceans of data, how can the data be utilized? A mechanical extension of our memory has been a dream since long. Vannavar Bush described in 1945 the MemEx machine in his famous paper *As We May Think*. In the 1980ies the hypertext was thought to do the job. The web in the 1990ies was an implementation of hypertext on a global scale. Linked data and the semantic web followed without really solving the problem.

The last decade we have been told to avoid information islands and the slogan has been "Open the data silos". Is it easier to find a needle in an enormous haystack than in many small? If we are satisfied with the result lists of the google-type answer, it is a clear yes. If we want to build scientific datasets which may be aggregated into larger datasets, we need common authority systems and we need to impose some common structure on the data. To do this in a meaningful way, we have to do an ontological analysis of why and how data is produced in our disciplines. That is, we need to understand our data and establish consistent and well-founded data models or ontologies. On the basis of those we can see how our data may be mapped to a common model for integration. Well defined data models are necessary to define standards for storage formats and may help us to write the necessary specification for contract excavators.

In the CAA context the main focus will and should be on innovative ICT-applications and good practice. The methodology of common consistent but flexible models for data integration will be a relatively small, but important core activity. The data and the artefacts is all what remains from an excavation. They must be handled with care. We need to create accept among the stakeholders that data are at least as important as the artefacts and need long term curation. This is a task for the entire CAA community as well as for the cultural heritage sector as a whole.