Avoiding a digital Dark Age

Advances in information and telecommunications technology present opportunities and risks for research and research data. These advances are propelling us into a new age of research. The question is: is this a golden age or a dark age for research information?

Researchers now use ever-increasing volumes of data about our world. Ever more common and powerful digital instruments and sensors churn out more data in a single session than a human being could deal with in a whole lifetime. By some projections, in 2018, "there will be more data being generated [annually] than has ever been generated in human history up to 2006".

Fortunately, the same ICT revolution has helped to deal with this tsunami of research data by empowering researchers to analyse, assemble and share research data. This is largely due to the emergence of data management systems, high-performance computing to manipulate large volumes of information, and infrastructure and protocols to network or federate information.

If all works out nicely, this will be a golden age for researchers: unlimited new online collections of data and research information with powerful tools for aggregating, analysing and accessing that information. But what are the risks?

Being able to preserve digital data is a must for a golden age of research information, and a major risk is therefore the rapid obsolescence of digital objects. File formats, software and hardware are constantly being superseded, so the curation of digital objects involves regularly migrating files into currently supported formats. Who will do this for important research information long after the original research group has been disbanded? Memory institutions such as libraries and archives will have a role, but research disciplines must also assist in identifying the intrinsic qualities that need to be preserved during migration.

Important research collections need to be under the stewardship of a sustainable body committed to – and able to ensure – the continuity of access to these digital research assets. Otherwise, these online research collections and data sets will never last long enough to revolutionise the way we do research. At worst, a new digital dark age will follow, where access to the previous generations’ information is severely compromised.

New research builds on previous research. In the new golden age, references to previous research or supporting data can include the actual digital objects or a link to the referenced digital object. Other scholars, in turn, refer to this research or include it in their digital works, and the new golden age builds on itself.

However, the mesh of information needs to be reliably persistent for future scholars to retrace these cross-collection workflows. The risk of a dark age occurs if the whole information infrastructure for scholarly communications is not permanent enough. The simple URL ‘address’ of the World Wide Web is insufficient. We need to use better systems of persistent identification to cope with changes of address; otherwise broken links will usher in another dark age for information.

The new model of scholarly information is decentralised with an unlimited number of online research collections hosted in various types of repositories, data centres and custom web applications. The beauty of this internet model is the organic growth of content by authors distributed all over the world. Combining this information into international grids is potentially part of the new golden age of research data and scholarly information.

The risk involved with this decentralised, distributed production model is the necessary diversity of systems and formats for storing the research data. Attempts to combine atmospheric data from around the globe can be stymied if the underlying data models are not compatible, so the tantalising possibility of aggregating our data remains elusive.

The patient development and disciplined application of community standards is the key to ensuring the golden age of mutual intelligibility does not turn into a dark age of tribal confusion.

The promised golden age includes sophisticated public services for search, discovery, access, analysis, visualisation, fusion, submission and presentation of research. For this to work we need intelligent data; the raw data needs to be structured, described, and ‘marked up’ with meta-information. This applies to scholarly literature as much as to data sets, because even text files need to be structured and marked up with discipline-specific meta-tags to participate in sophisticated bibliographic, data-mining and discovery services.

In the dark age of information we bequeath to our sons and daughters an unending sea of ones and zeros with no standard structure, description or provenance data. Extracting useful information from this ‘dumb data’ will be a time-consuming process. The golden age is predicated on openness, a willingness to grant access to scholarly outputs and research data. With the advent of the World Wide Web as a core part of popular culture, there is a new expectation that everything should be findable and accessible online. And commonly available software empowers authors and data scientists to self-publish their work.

Copyright and digital rights management are not necessarily risks to this openness. The risks lie rather with the general ignorance of the rights and responsibilities in this area or with the lack (or non-adoption) of clear protocols for expressing these rights.

Openness of research data has social barriers in some disciplines where primacy and sole use of data is important to academic reputation. Other disciplines have adopted, at a community level, a greater expectation of immediate open access to research data.

Advances in ICT technology are enabling the prospect of a golden age of research information. However, the barbarians are massing outside the empire and unless we invest to secure digital longevity, persistent identification, interoperability, richness of data and open access, a regression into a digital dark age is also possible.