Long Term Preservation of Digital Content: Issues and Approaches

Òåçèñû â ôîðìàòå RTF (41.82 kB 2008-07-15 16:26:14)
Òåçèñ â ôîðìàòå PDF (77 kB 2008-07-15 16:26:39)

Introduction

Digital preservation is an issue that affects every citizen in the
information society. It covers a whole spectrum of issues, from long-term
access to and use of personal digital objects to the complex area of
information objects’ lifecycle management in big institutions from the
engineering, governmental, research and cultural heritage sectors.
E-government, e science and e-culture greatly depend on proper storage and
access to huge collections of digital resources, which should not be affected
by changes in the technological environment.

The importance of this area had been recognised
by the European Commission, which launched the Digital Library Initiative as one of its four
flagship initiatives in the i2010 programme.[1]
Digitalization, accessibility online and digital preservation are the basic
areas of work in the digital libraries domain. From these three areas, digital
preservation has a special place because it guarantees the
interoperability of the digital resources in the future.

Despite its rapidly
increasing role, Digital Preservation area has not yet reached a level of
maturity similar to its constituent research domains. It is a multidisciplinary
area involving researchers and practitioners from several fields ranging from
Information Retrieval to Library and Archival Science, Content Management,
Modelling, Simulation, Human-Computer Interaction, Scholarly Communication and
Natural Language Processing. In particular, an issue which needs to be
addressed is automation in digital preservation, as identified by the DigitalPreservationEurope’s research roadmap published
in the second half of 2007.[2]
It highlighted the urgent need for accelerated development in automating
digital repository processes.

Automation in Digital Preservation

The field of digital preservation has been
developing at an increasing pace in recent years. However, digital preservation
is still not exploiting the potential of automated methods. One of the
fundamental areas in digital preservation automation is information object
lifecycle management, which includes extraction and handling of preservation
metadata.

In the preservation metadata field much
of the work has concentrated on modelling metadata schemas, paying
comparatively little attention to the development and integration of automated
extraction and management tools.[3]
Current research in automated metadata extraction, on the other hand, has been
dealing with metadata in the general sense, not with metadata designed to
support preservation.

The current practice reflects several
influences. On the one hand, we face an increasing constriction of the
“metadata bottleneck”:[4]
it is a cause for alarm that a growing amount of born-digital and digitally
reformatted material does not have any metadata attached at all (see, for
example, the research of Zhang and Iastram, who found that in a sample of 2400
websites only 62.83% contained embedded HTML metadata[5]). On the other hand, the poor quality of
metadata continues to be an obstacle. Studies show that the quality of
manually-created metadata depends heavily on the institutional framework,
personal motivation and competence,[6]
and metadata redundancy is emerging as an increasingly common problem.[7]
Redundancy appears in two ways: first when various institutions create metadata
for the same digital object (lack of coordination), and second when different
digital objects are given similar metadata (digital objects created with minor
variations). Without adequate metadata, management of digital entities is not
feasible, and the manual creation of such metadata is resource intensive. This
makes it obvious that automatic generation of metadata is an absolute must.

The scepticism towards the implementation of
automated preservation extraction methods may be caused by the information
retrieval results reported in research – precision between 0.79 and 0.96 and
recall between 0.62 and 0.99, with different values for different metadata
elements. Taking into account the importance of preservation metadata, and the
risks of working with digital objects which do not have any metadata at all,
even this seems to be a better option than waiting for the invention of the
method which would guarantee 100% accuracy.

To achieve progress and establish a common
framework, an analysis of the shortcomings of automated metadata creation
should be combined with a study of the possibilities for developing combined
approaches based on analysing manual versus automated extraction of different
elements and on adding intelligent elements to automated metadata extraction
methods, such as the analysis of the genre of the document to select the best
automated extraction tool, as well as implementing self-documenting components
into the metadata lifecycle to assist the process.

Automated metadata extraction is only one
automation issue, but there are also further applications which need to be
developed in order to automate overall digital object lifecycle management. The
discussion of possible application scenarios for various digital preservation
tasks will make it possible to address the multifaceted nature of digital
preservation.

Conclusion

The paper will discuss the following topics:

-
What is the real
place of automation in digital preservation?

-
How should we
better understand user needs? How adequate to the user needs are the current
approaches?

-
What are the
general and specific application scenarios in digital preservation?

-
What are the recent
achievements of DigitalPreservationEurope, Planets, CASPAR and SHAMAN projects
supported by the European Commission?

Special place will be given to the question of what preservation
elements should be considered by projects aimed at developing digital
resources. This is important in order to build awareness of the importance of
long-term preservation while planning the development of new resources in order
to enhance their sustainability.

Äëèòåëüíîå ñîõðàíåíèå öèôðîâîãî êîíòåíòà:
ïðîáëåìû è ïîäõîäû

Ì. Ï. Äîáðåâà

Óíèâåðñèòåò Ñòðàòêëàéä, Ãëàçãî, Øîòëàíäèÿ, Âåëèêîáðèòàíöèÿ

Èíñòèòóò ìàòåìàòèêè è èíôîðìàòèêè Áîëãàðñêîé ÀÍ, Ñîôèÿ, Áîëãàðèÿ

Äîëãîñðî÷íîå õðàíåíèå íàó÷íûõ è êóëüòóðíûõ öåííîñòåé
ÿâëÿåòñÿ âàæíîé ïðåäïîñûëêîé óñïåøíîãî ðàçâèòèÿ èíôîðìàöèîííîãî
îáùåñòâà. Îíî íâÿçàíî íå òîëüêî ñ ôèçè÷åñêîé ñîõðàííîñòüþ äàííûõ, íî è ñ
ãàðàíòèåé âîçìîæíîñòè èõ
èñïîëüçîâàíèÿ ïðîãðàììíûì îáåñïå÷åíèåì áóäóùåãî. Ïðè ýòîì îñîáî âàæíóþ ðîëü
èìåþò ïîíÿòèÿ èíòåãðàöèè, àóòåíòè÷íîñòè,
äîñòîâåðíîñòè äàííûõ. Â ðàáîòå ðàññìîòðåíû îñíîâíûå
íàïðàâëåíèÿ ðàáîò â îáëàñòè äîëãîñðî÷íîãî õðàíåíèÿ ýëåêòðîííûé
ðåñóðñîâ íà ïðèìåðàõ åâðîïåéñêèõ ïðîåêòîâ CASPAR, DigitalPreservationEurope, Planets è SHAMAN.

[1] European
Union. i2010: Digital Libraries Initiative. URL:
http://ec.europa.eu/information_society/activities/digital_libraries/index_en.htm.

[2] DPE: Digital
Preservation Europe. EC Contract No. IST-2005-34762. [Online]. URL: http://www.digitalpreservationeurope.eu.

[3] Lavoie, B.,
Gartner R.: Preservation Metadata. A joint report of OCLC, Oxford Library
Services, and the   Digital Preservation Coalition (DPC), published

electronically as a DPC Technology Watch Report (No. 05- 01). URL: http://www.dpconline.org/docs/reports/dpctw05-01.pdf
(2005).

[4] Liddy, E.D.: A Breadth of NLP
Applications. IN: ELSENEWS of the European Network in Human Language
Technologies. Winter. (2002).

[5] Zhang, J. Jastram, I.: A Study of
the Metadata Creation Behavior of Different User Groups on the Internet.
Information Processing & Management, Vol. 42, pp. 1099–1122. (2006).

[6] Crystal A., Greenberg J.: Usability
of a Metadata Creation Application for Resource Authors, Library &
Information Science Research V. 27(2), pp. 177–189 (2005).

[7] Foulonneau M.: Information
Redundancy across Metadata Collections. Information Processing &
Management. Vol. 43(3), Special Issue on Heterogeneous and Distributed IR, pp.
740–751 (2007).