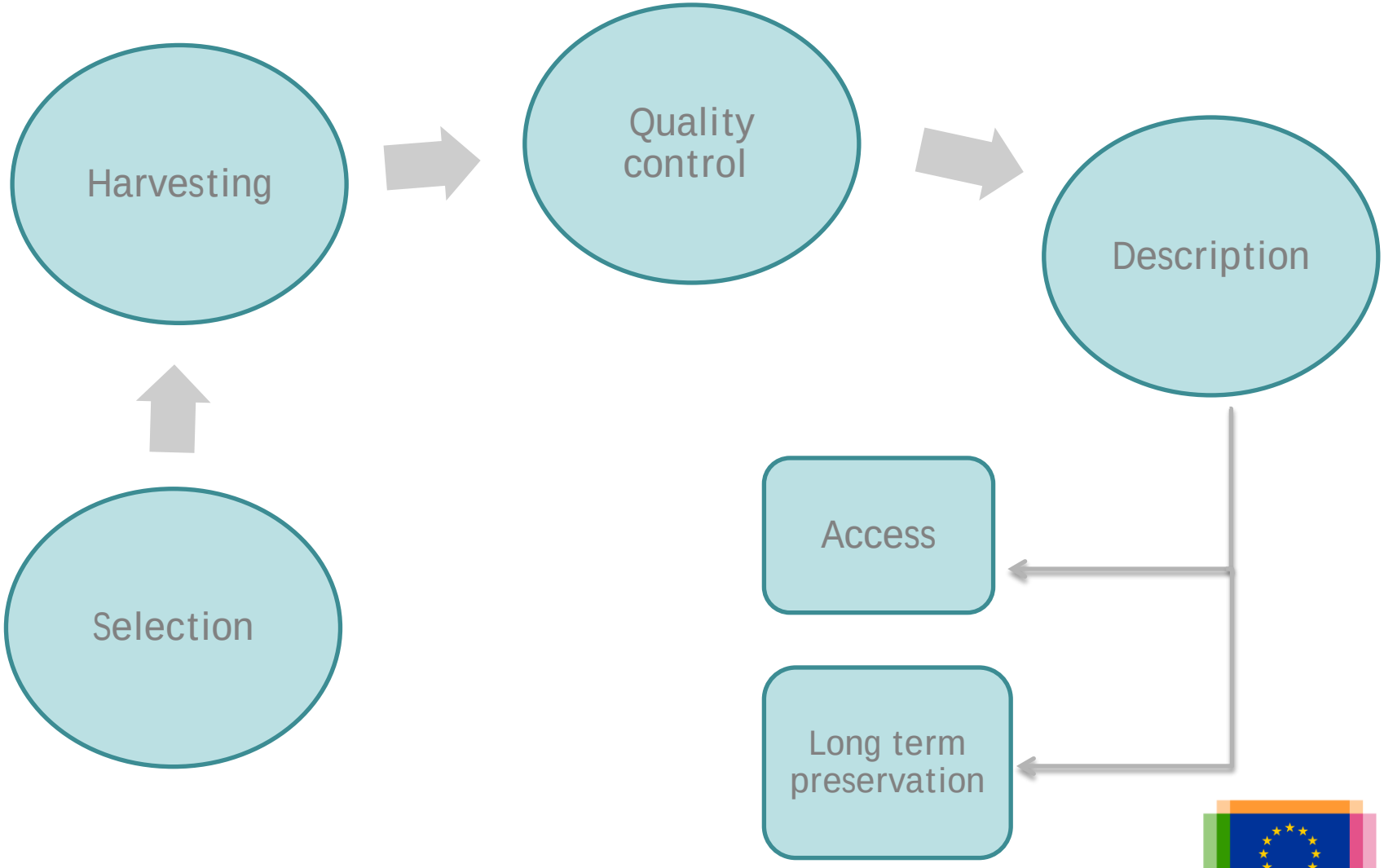
Creating a web archive for the EU institutions' websites

Achievements and challenges

Els Breedstraet
Coordinator web preservation
Publications Office of the EU

IIPC Web Archiving Conference, London, 16 June 2016

Creating the archive



Selection

	**Internal: Publications Office / Historical Archives of the EU**
	• **Seed list of ca 70 europa.eu URLs** • **Covers mainly sites of EU institutions, bodies and agencies**

	• **Selection policy**
	  • **To be limited to europa.eu domain and subdomains?**
	  • **Which seeds to add? Which selection criteria?**
	  • **Who decides to add or not?**
	  => need for policies + curator and/or selection board
	• **Seed management:**
	  • **How to propose new seeds => need for selection tool**
	  • **How to keep list up to date?**

Harvesting

	Outsourced (Internet Memory Research)
	• Quarterly crawling of seed list • Using Heritrix • Runs successful since end of 2013 • Till mid 2016: ARC, afterwards WARC

	• Dynamic content
	• Social media
	• Multilingual harvesting
	• Ad hoc demands/campaigns
	  • How to decide to do or not?
	  • Planning and budgeting

Quality control

	**Automatic: outsourced** **Manual: website owners, coordinated by the Publications Office**
	• **Automatic checks + manual / visual control** • **Network of quality controllers** • **Structured follow up of feedback => long term improvement harvests**

	• **Automatic checks: more? Reporting?**
	• **Manual control**
	  • **Need for quality control tool**
	  • **Need for crawl reports**
	• **Awareness raising -> webmasters to be encouraged to make "well preservable" sites**

Description / metadata

	**Automatic (up to now)**
	• **Automatic capturing of descriptive and technical MD at the source** • **Proposal for minimal set of MD (descriptive, technical and provenance) ready**
	• **Awareness raising -> webmasters to be encouraged to add more and/or more structured MD at the source** • **Provenance MD** • **Post-crawl manual description?** • **How to exploit MD to improve access?**

Access

	**Outsourced (Internet Memory Research)**
	• **Fully open archive:** http://collections.internetmemory.org/haeu • **Basic search and browse options**
	• **Audience / use cases?** • **Usage statistics** • **Branding / customisation** • **Data protection, copyright, etc.** • **Multilingual access** • **Promotion and communication** • **Centralised access to EU related web archives? Cooperation?**

Long term preservation

	**Internal: digital repository at Publications Office**
	• **First tests started with ingestion in internal trusted long term digital preservation repository**
	• **Adaptation information architecture** • **Provenance metadata** • **Trustworthy Digital Repository Certification (ISO16363)**

Conclusions

- Solid basis for an EU institutional web archive is in place, accessible to everyone
- Many challenges, a lot of work ahead
- Many dreams and ideas for the future
  - One step at the time
  - Cooperation?

More information

Have a look at our web archive:
http://collections.internetmemory.org/haeu

Questions or suggestions? Contact us:
op-web-preservation@publications.europa.eu

Thanks for your attention!

Questions?