The importance of blogs as a distinct class of Web resource has received considerable attention in recent years, notably at iPRES. The need to capture this dynamic, cumulative content for future access has been recognised by many institutions and projects, and a variety of tools and approaches have emerged.

This poster presents a summary of key results of interest from a comparative analysis of three distinct to collections of archived blog content that vary considerably in their facilities and approach. These are:

- The BlogForever project, funded by the European Union, has developed an integrated platform, comprising a harvesting methodology and associated content management system, for creation, management and preservation of blog collections.
- The London School of Economics (LSE) preserves its academic blogs by creating and depositing PDF renditions of blog posts into an existing Institutional Repository.
- The UK Web Archive, operated by the British Library, which collects and preserves blog content from the UK Blogosphere. This collection represents a cross section of UK Web blogs containing a wealth of material which will be of value to researchers now and in the future.

The study considered each archive’s performance against key aspects of digital preservation based on established standards and practice, such as PREMIS, OAIS and IIPC, and the expectations of user communities, for example:

1. Long Term Preservation (OAIS): does the repository offer sufficient control of the content to ensure long-term preservation?
2. Designated Community: does the repository identify a Designated Community that should be able to understand the information provided; and is the content independently understandable and available to the Designated Community?
3. Preservation metadata: does the repository support the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context?
4. Metadata encoding and transmission (METS): is there metadata necessary for both the management of digital objects within a repository and exchange of such objects between repositories (or between repositories and users)?
5. Completeness: is the collection underpinned by a sound selection policy to ensure comprehensive coverage? (IIPC Selection for Web Archives)?

To ensure consistency of comparison across the platforms, we defined a small set of interesting and exemplary blogs, each of which was available for comparison in at least two of the platforms being studied. Some of our conclusions are illustrated in the panels on the right. Other conclusions include:

- That preserving parsed blog content (BlogForever) offers greater benefits in terms of discovery and fine-grained retrieval than preserving entire crawled websites (as per UK Web Archive).
- That websites stored in the WARC format (UK Web Archive) are more robust and better supported as coherent, preservable digital entities.
- XML-based blog content, capable of being exported into numerous library and metadata formats such as MARC XML, Dublin Core and METS, offers more flexibility for interoperability and sharing than WARC.
- That PDF renditions of blogs (LSE) are easier and quicker to produce than using traditional web-archiving methods, but may in turn introduce additional preservation challenges.
- That renditions of blog content viewed through the Wayback Machine (UK Web Archive) are perceived as more complete with regards to look and feel, attachments and layout than pre-processed renditions stored in XML (BlogForever).
- That research value to scholars is enhanced by maintaining and indexing an aggregated collection of micro-detail from the blogosphere (authors, tags, comments).
- Aggregated collection of textual blog content will potentially be extremely useful to text-mining projects that are concerned with finding particular types of patterns, e.g. the evolution of language used on the internet, that cannot be easily discerned through the more usual title-based approach.

The study included:

- A small set of interesting and exemplary blogs, each of which was available for comparison in at least two of the platforms being studied.
- Some of our conclusions are illustrated in the panels on the right. Other available for comparison in at least two of the platforms being studied.
- A variety of tools and approaches have emerged.

We selected a small sample of blogs whose live sites were still available online, and compared the number of live blog posts with the number of posts in each archive.

Results: the Heritrix-based approach delivers an almost complete gather of posts on a blog. Both the other approaches involve appraisal and selection of blogs and posts for inclusion, which contributes to the lower proportion of blog posts gathered. RSS-based harvesting also has its limitations.

We analysed the representation of blog posts in the archive and created a simple metric based on the explicit availability and visibility of core metadata for search and discovery: author, date of publication, post title and blog title.

Results: The Invenio and EPrints-based approaches proved much more effective at providing and exposing the metadata. Since Heritrix and WARC render a website as a string of encoded page requests, it is not possible to include metadata specific to blog posts and pages.

We measured the extent and means by which the archive produces interoperable catalogue or descriptive metadata, discoverable by a protocol standard and use of standard storage formats for potential data exchange. Standards tested for include DC, METS, MARC, BibTEX, OAI-PMH, SRU.

Results: the Invenio and EPrints-based platforms provided built-in and extensible support for many key interoperability standards.