



The National Archives

Towards the development of a test corpus of digital objects for the evaluation of file format identification tools and signatures

Ross Spencer

7 December 2011

1. Why?

- Improved trustworthiness and validation
- Encourage tool development
- Encourage collaboration and dialogue within community

2. Existing solutions?

The screenshot shows the Digital Corpora website with a main article titled "Govdocs1" dated September 6th, 2010. The article discusses the creation of a corpus of 1 million freely-redistributable files for forensic research. It explains that these files were obtained by searching for random words and numbers in the .gov domain. The article lists several ways the corpus is available: as 1000 directories of 1000 files each, as 1000 ZIP files, or as a set of 10 subsets of 100,000 files each. A search interface for the corpus is visible at the bottom of the article.

The screenshot shows the ConversionSoftwareRegistry website. It features a search bar with options for "Name", "Extension", and "Search". Below the search bar, it displays search results for "govdocs1", showing "Number of extensions: 1775, MIME: 488, PUID: 415, UTI: 305". There are also links for "Search Extension, MIME, Promom ID (PUID) or Apple UTL".

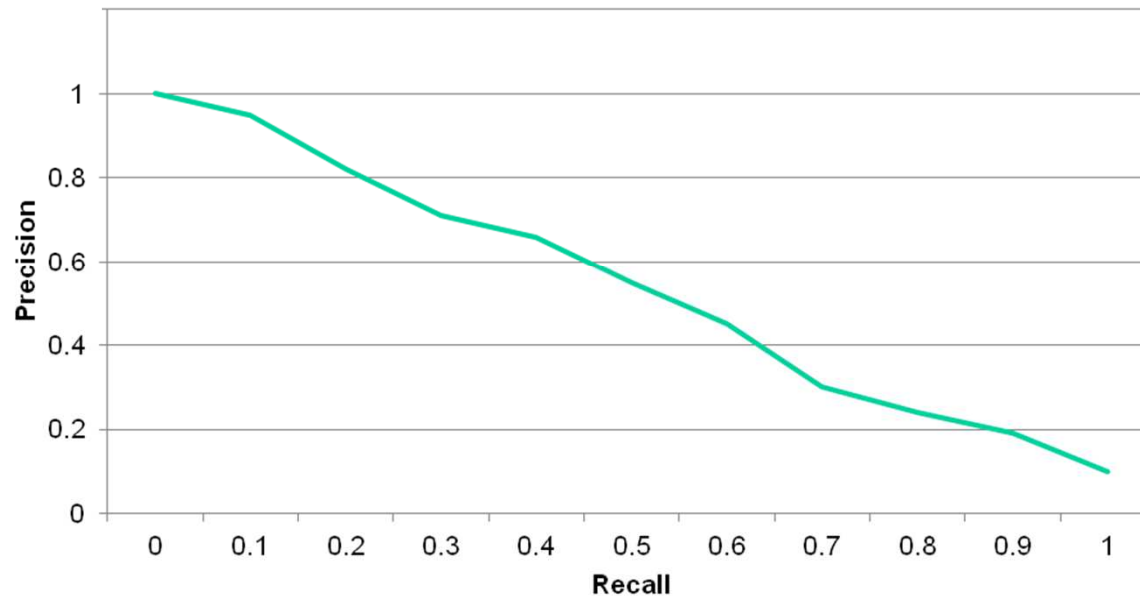
The screenshot shows the website for the University of Waterloo's Image Repository, part of the fractal coding and analysis group. The page lists various image sets available for download, including "Greyscale Set 1" (12 small greyscale images), "Greyscale Set 2" (12 medium greyscale images), and "Color Set" (9 large full color images). It also lists "Some External Links" such as "The International Conference on Image Analysis and Recognition" and "USC Image Database". A grid of image thumbnails is visible under the "Greyscale Set 1" section.

3. Practicalities

- Provenance
- Persistency
- Usability
- Security
- Maintenance
- IPR and copyright
- Access conditions

4. Metrics

Precision/Recall Analysis



5. Conclusion

- Broadly covered areas we need to consider to set up a test corpus
- Potential to provide quality assurance on digital preservation tool sets
- Potential to provide greater confidence and transparency
- Potential to become a collaborative platform in the community
- Not to be taken lightly. Planning, set-up, maintenance, support and use

Thank You