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

Ross Spencer

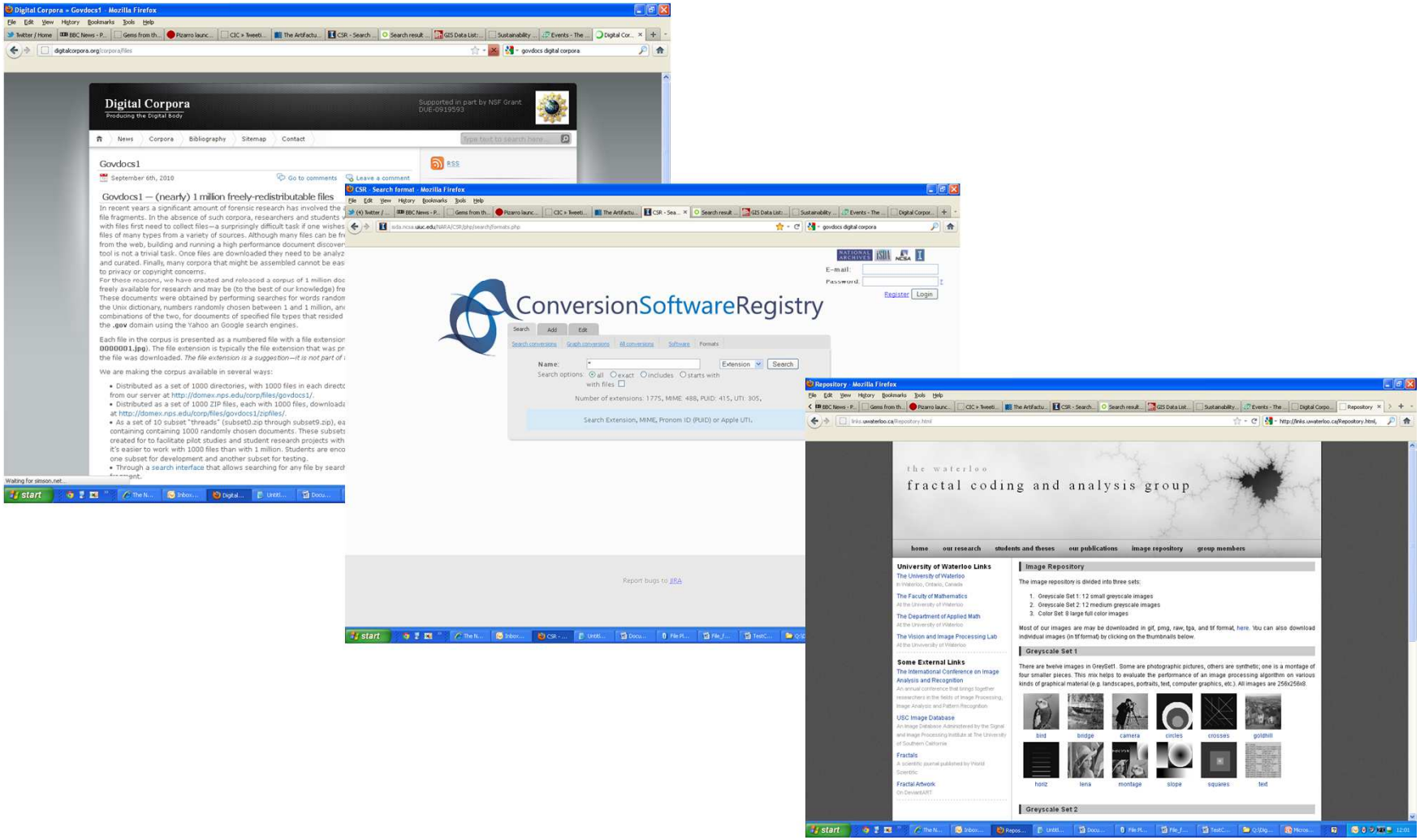
7 December 2011

The National Archives

1. Why?

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

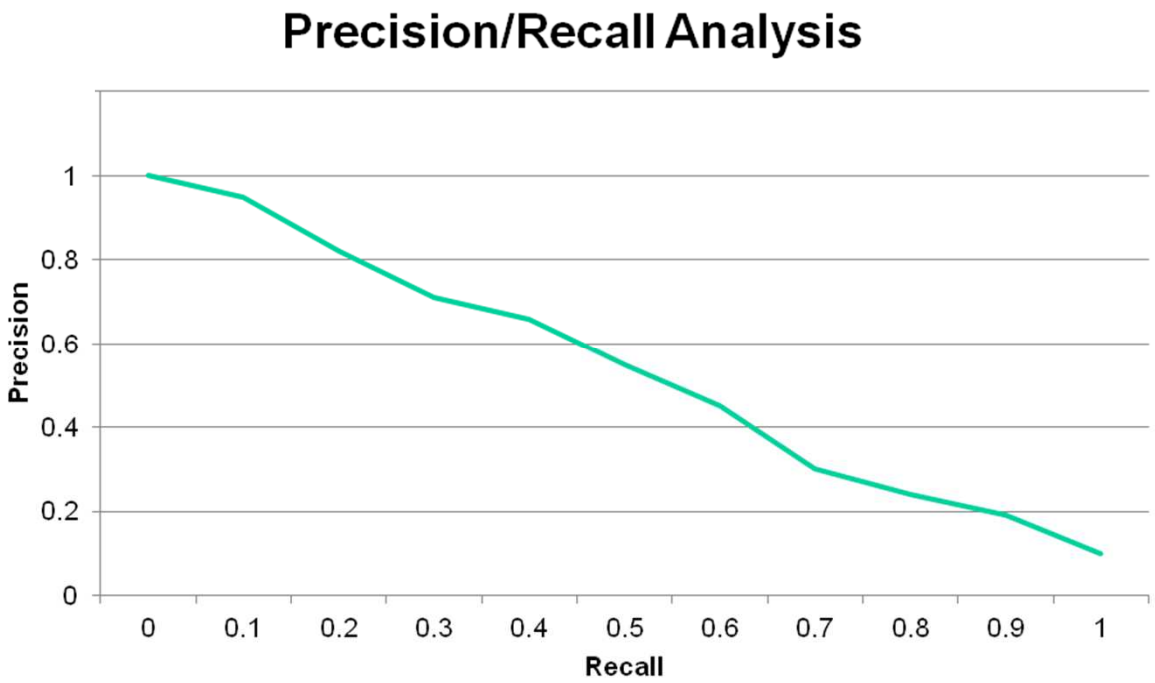
2. Existing solutions?



3. Practicalities

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

4. Metrics


Precision/Recall Analysis
Precision
1
0.8
0.6
0.4
0.2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Recall

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