THE AUSTRALIAN NATIONAL UNIVERSITY

The digital scholar's workbench

Ian Barnes

ELPUB 2007 Vienna — 13th to 15th June 2007

This work was supported by the Australian government through:

Preservation of text

This is a story in three parts, each concerned with a question about text preservation:

1. What format should we use?
2. How do we convert documents into that format?
3. How do we get authors to actually do this?

What format?

- Word? PDF? ODF? XML??
- Criteria:
  - Structure vs appearance
  - Open, free standards-based vs proprietary, closed
  - Based on plain text vs binary
  - Easy to transform/migrate/process
- On these criteria, only XML is any good, but what XML?
  - DocBook? TEI?
  - XHTML + …
  - Custom format?

How to convert into XML?

- This is a technical question
- It can be difficult — word processing formats are a big mess
- The problem is mostly solved if authors use styles from a good template (e.g. the ICE template from University of Southern Queensland)
- Without styles, this is a work in progress

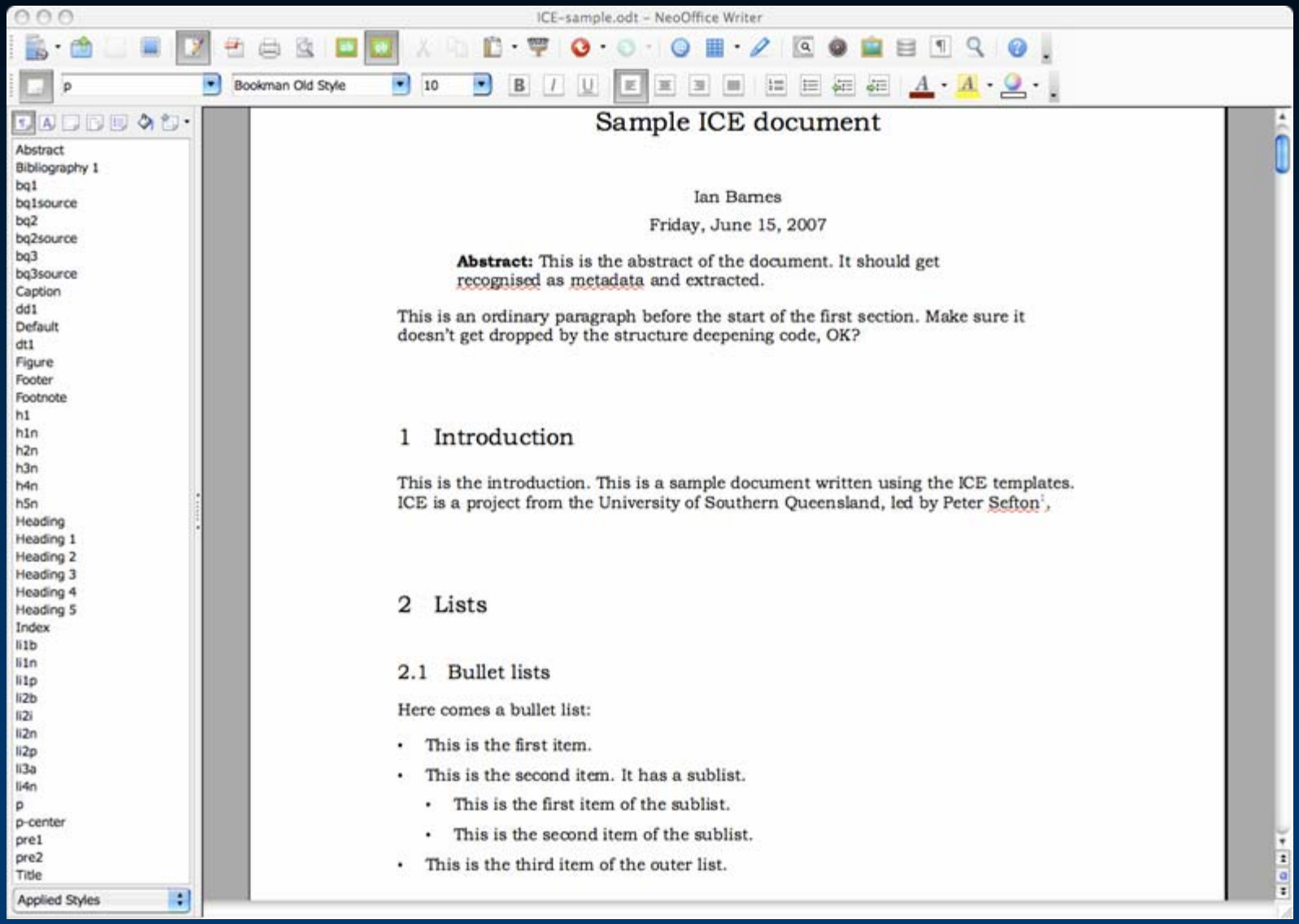
How do we get people to *do* this?

- This is *not* a technical question
- Low deposit rate is a big problem for repositories
- Why?
  - People don't care (until age 64)
  - It's too much work

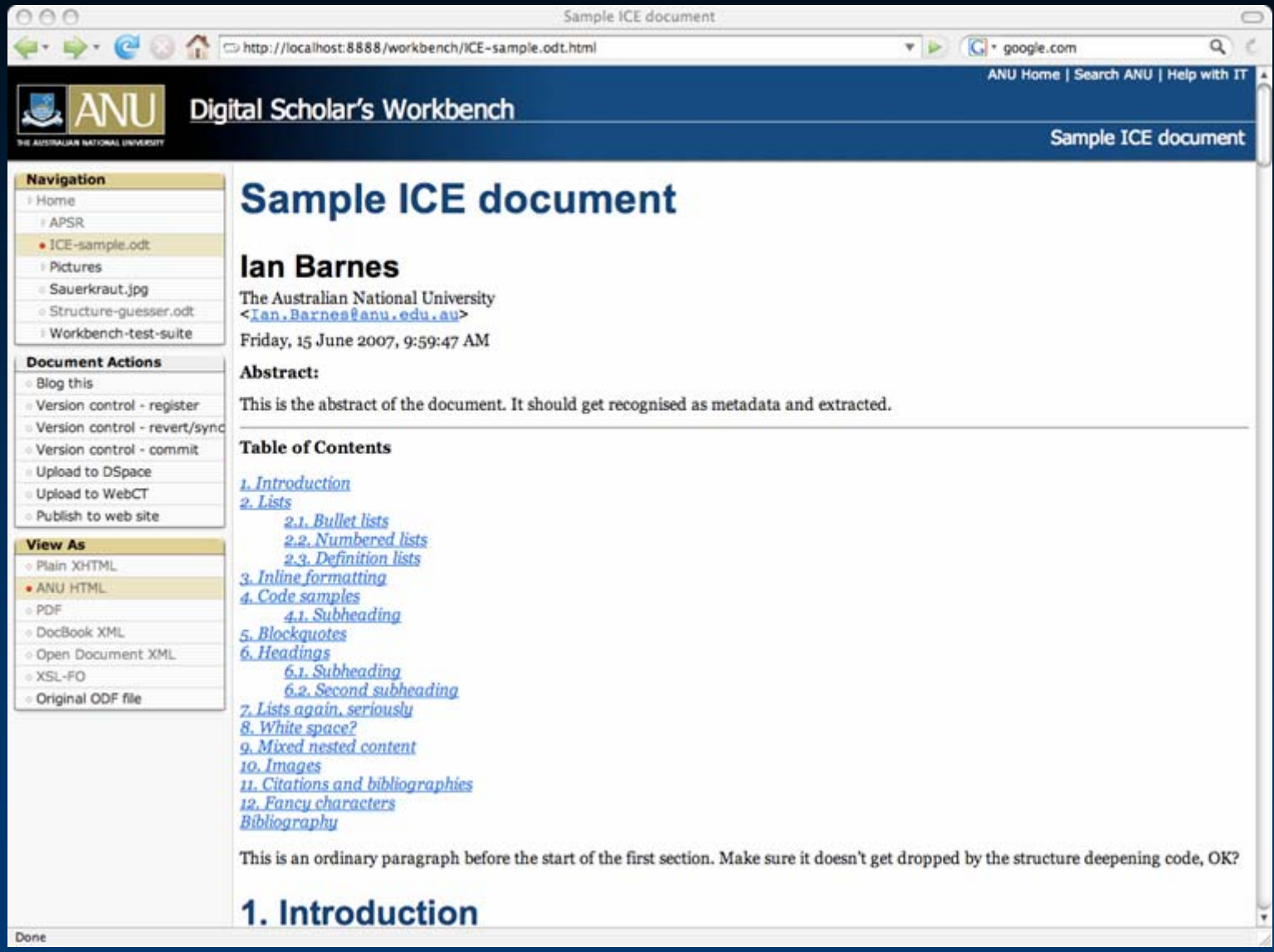
The solution: offer more, make it worthwhile

- Multiple publishing pathways
- Instant feedback/turnaround
- Interoperability
- … and much more …

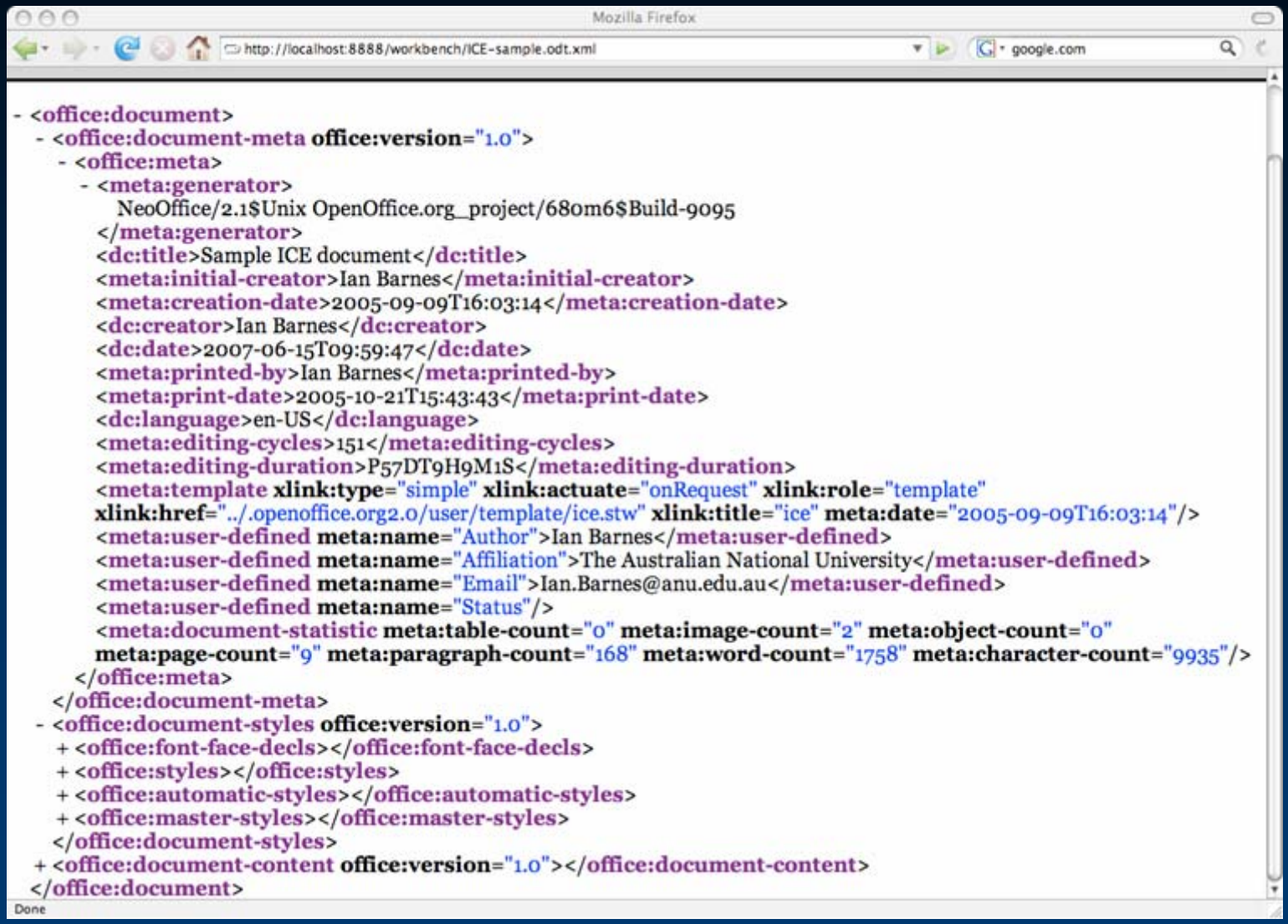
Document in word processor



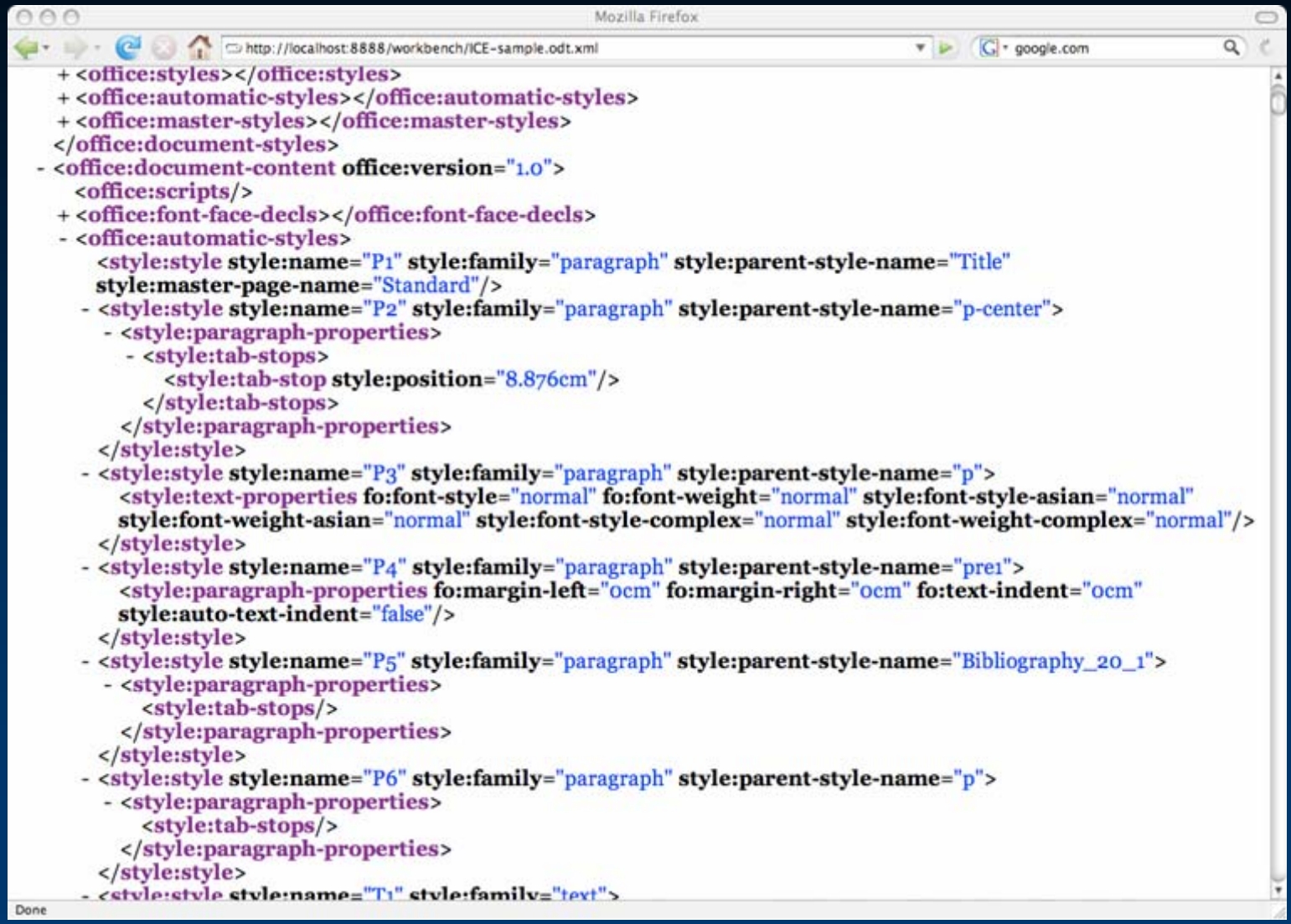
Converted automatically to HTML



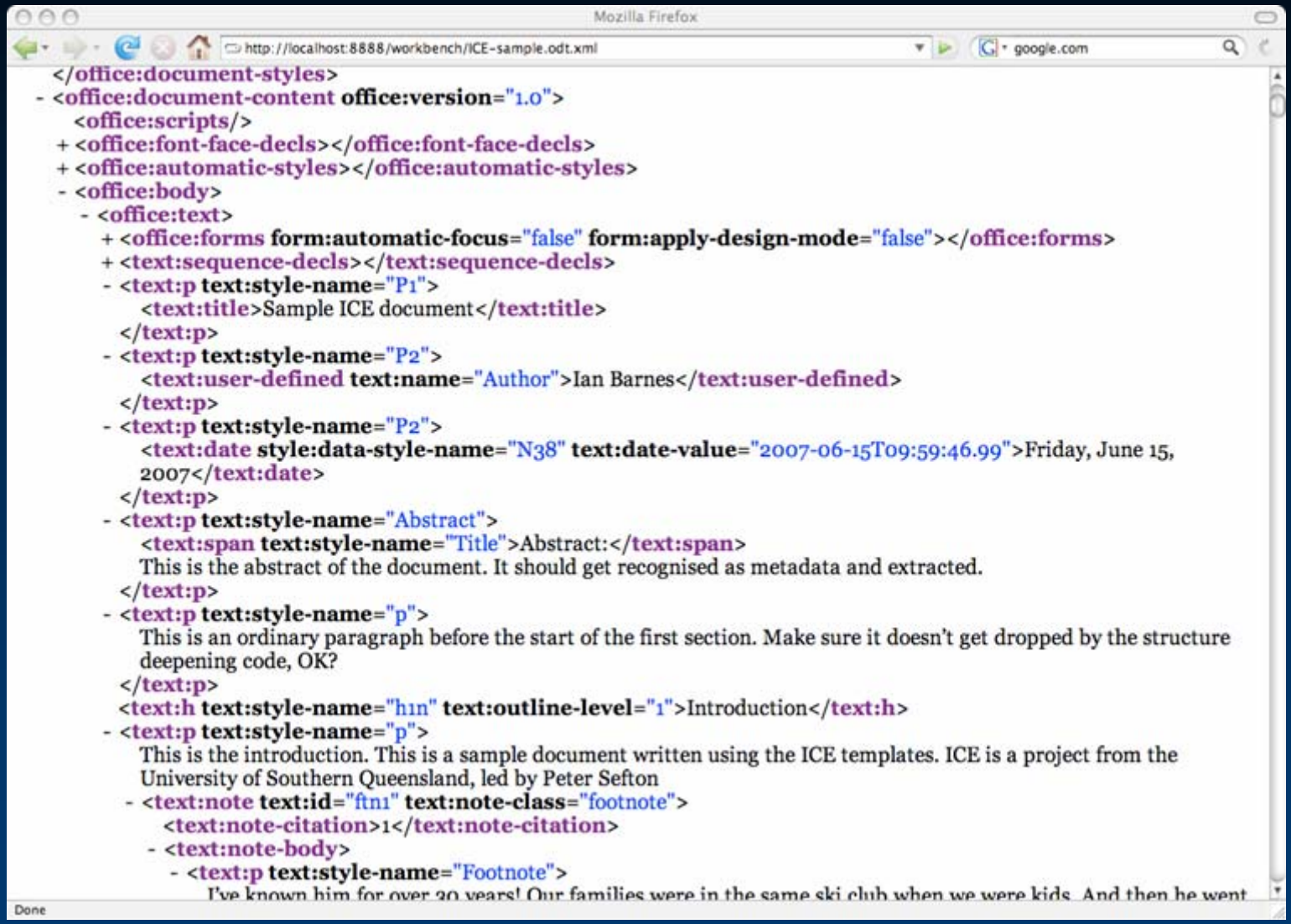
Open Document Format XML



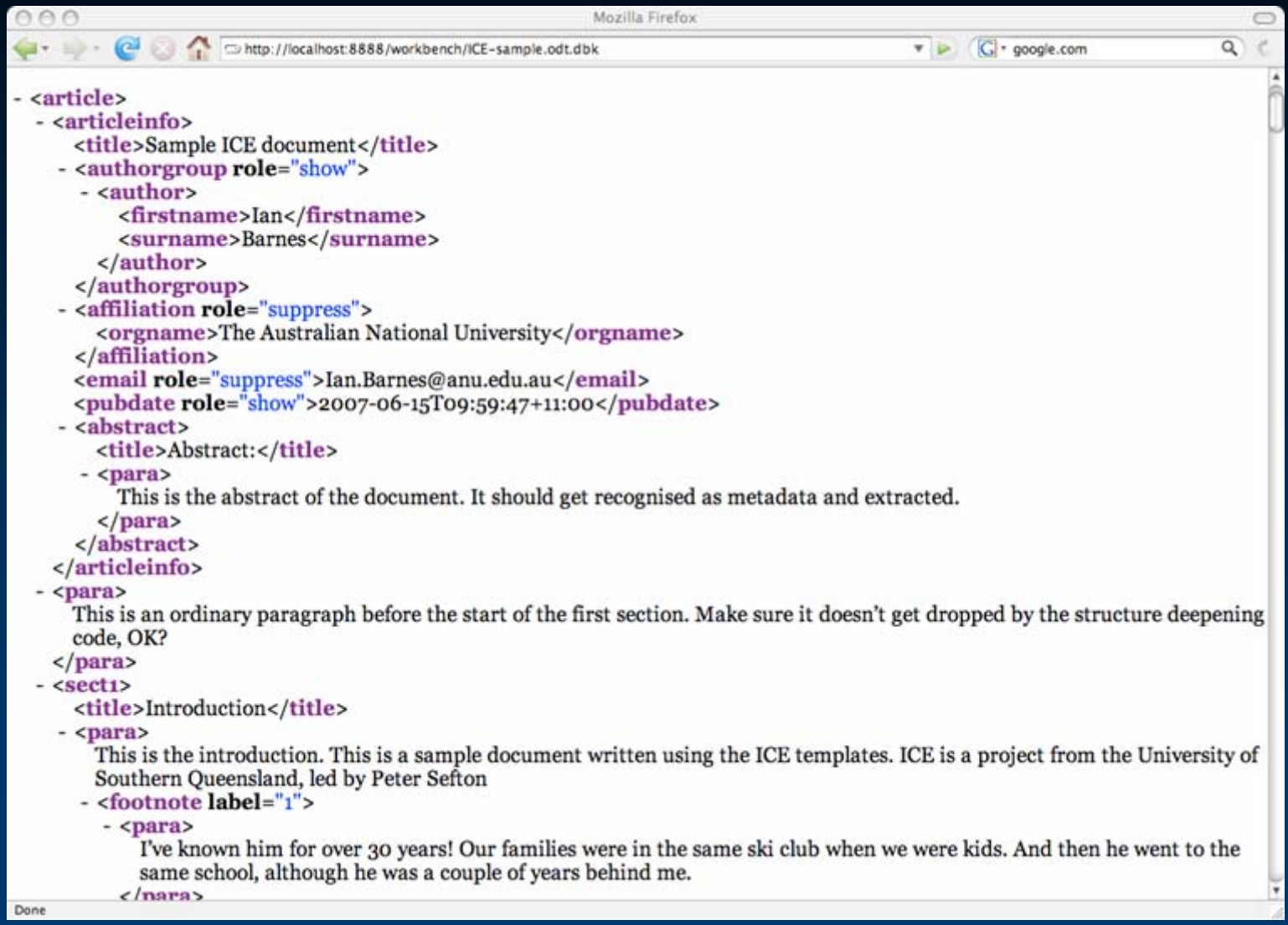
Open Document Format XML



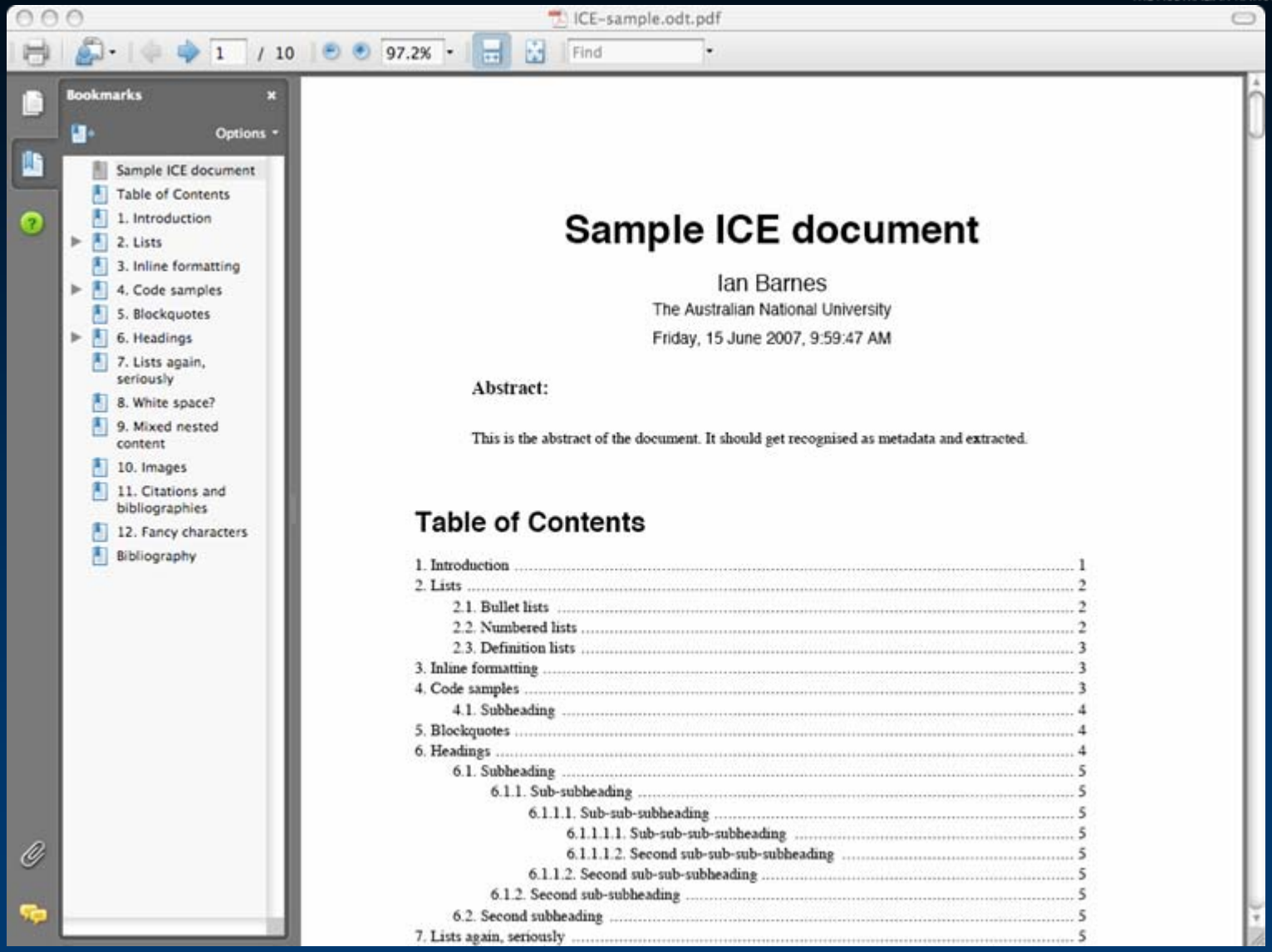
Open Document Format XML



DocBook XML


Mozilla Firefox
http://localhost:8888/workbench/ICE-sample.odt.dbk
google.com
- <article>
- <articleinfo>
<title>Sample ICE document</title>
- <authorgroup role="show">
- <author>
<firstname>Ian</firstname>
<surname>Barnes</surname>
</author>
</authorgroup>
- <affiliation role="suppress">
<orgname>The Australian National University</orgname>
</affiliation>
<email role="suppress">Ian.Barnes@anu.edu.au</email>
<pubdate role="show">2007-06-15T09:59:47+11:00</pubdate>
- <abstract>
<title>Abstract:</title>
- <para>
This is the abstract of the document. It should get recognised as metadata and extracted.
</para>
</abstract>
</articleinfo>
- <para>
This is an ordinary paragraph before the start of the first section. Make sure it doesn't get dropped by the structure deepening code, OK?
</para>
- <sect1>
<title>Introduction</title>
- <para>
This is the introduction. This is a sample document written using the ICE templates. ICE is a project from the University of Southern Queensland, led by Peter Sefton
- <footnote label="1">
- <para>
I've known him for over 30 years! Our families were in the same ski club when we were kids. And then he went to the same school, although he was a couple of years behind me.
</para>
Done

Automatically generated PDF



Proposed features

- One-click archiving including metadata extraction (already demonstrated with DSpace)
- Reformatting for journal/conference submission
- Publish to web site
- Publish to blog
- Complex and large documents (multi-part)
- Version control
- Collaboration/interoperability/round-tripping