



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

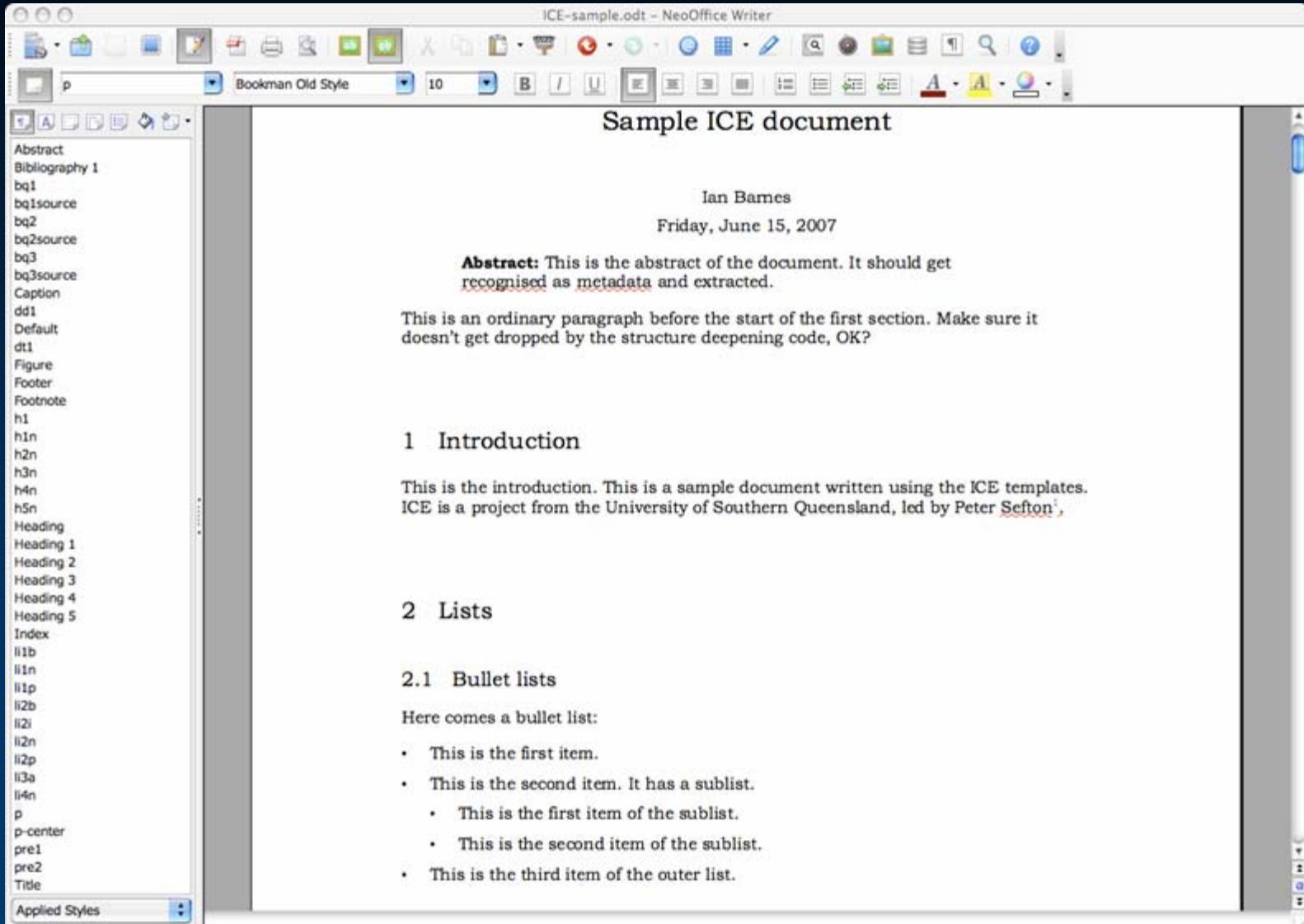
How do we get people to *do* this?

- 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



The screenshot shows the NeoOffice Writer interface. The title bar reads 'ICE-sample.odt - NeoOffice Writer'. The document content is as follows:

Sample ICE document

Ian Barnes
Friday, June 15, 2007

Abstract: This is the abstract of the document. It should get recognised as metadata and extracted.

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?

1 Introduction

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.

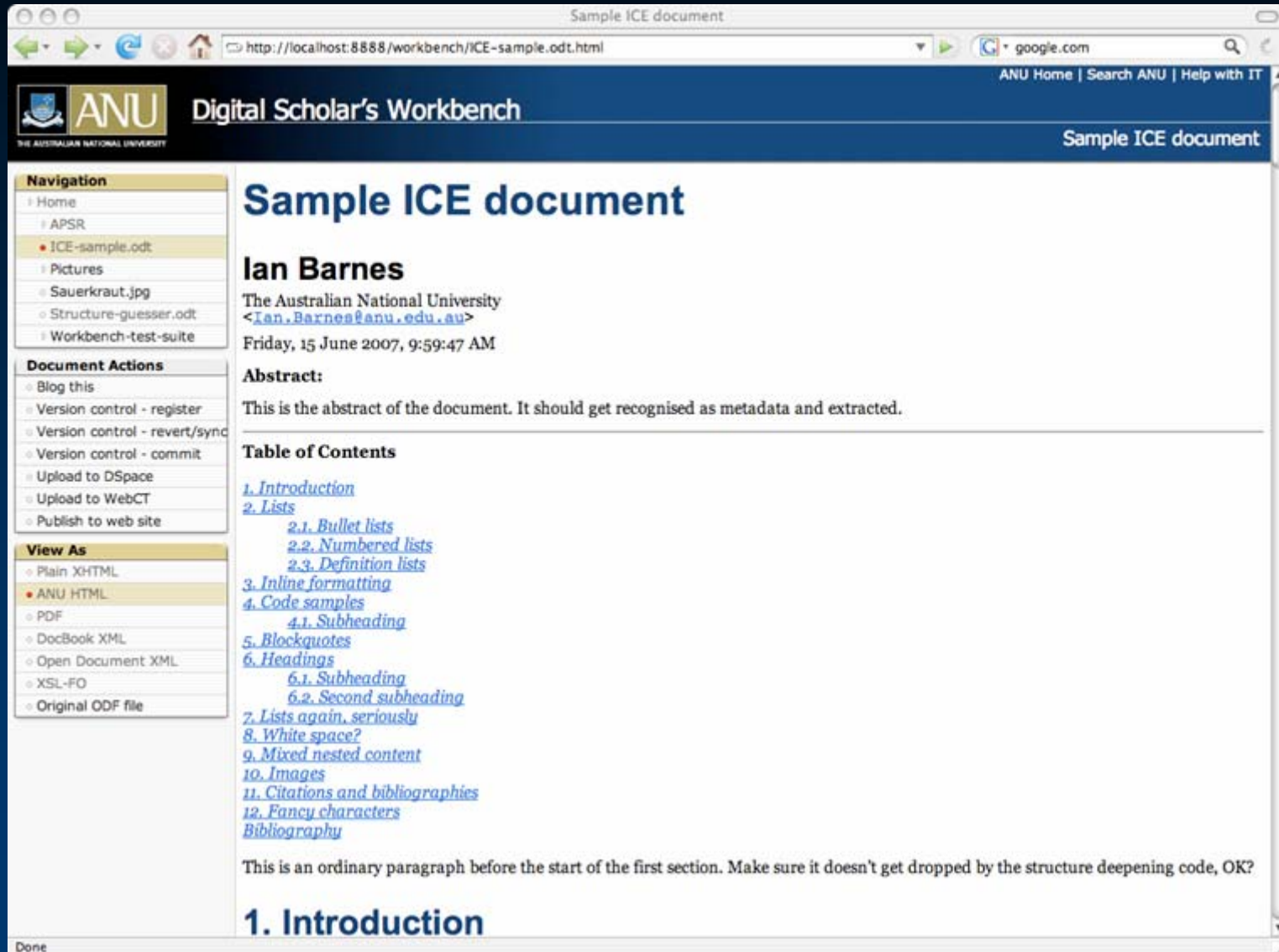
2 Lists

2.1 Bullet lists

Here comes a bullet list:

- This is the first item.
- This is the second item. It has a sublist.
 - This is the first item of the sublist.
 - This is the second item of the sublist.
- This is the third item of the outer list.

Converted automatically to HTML



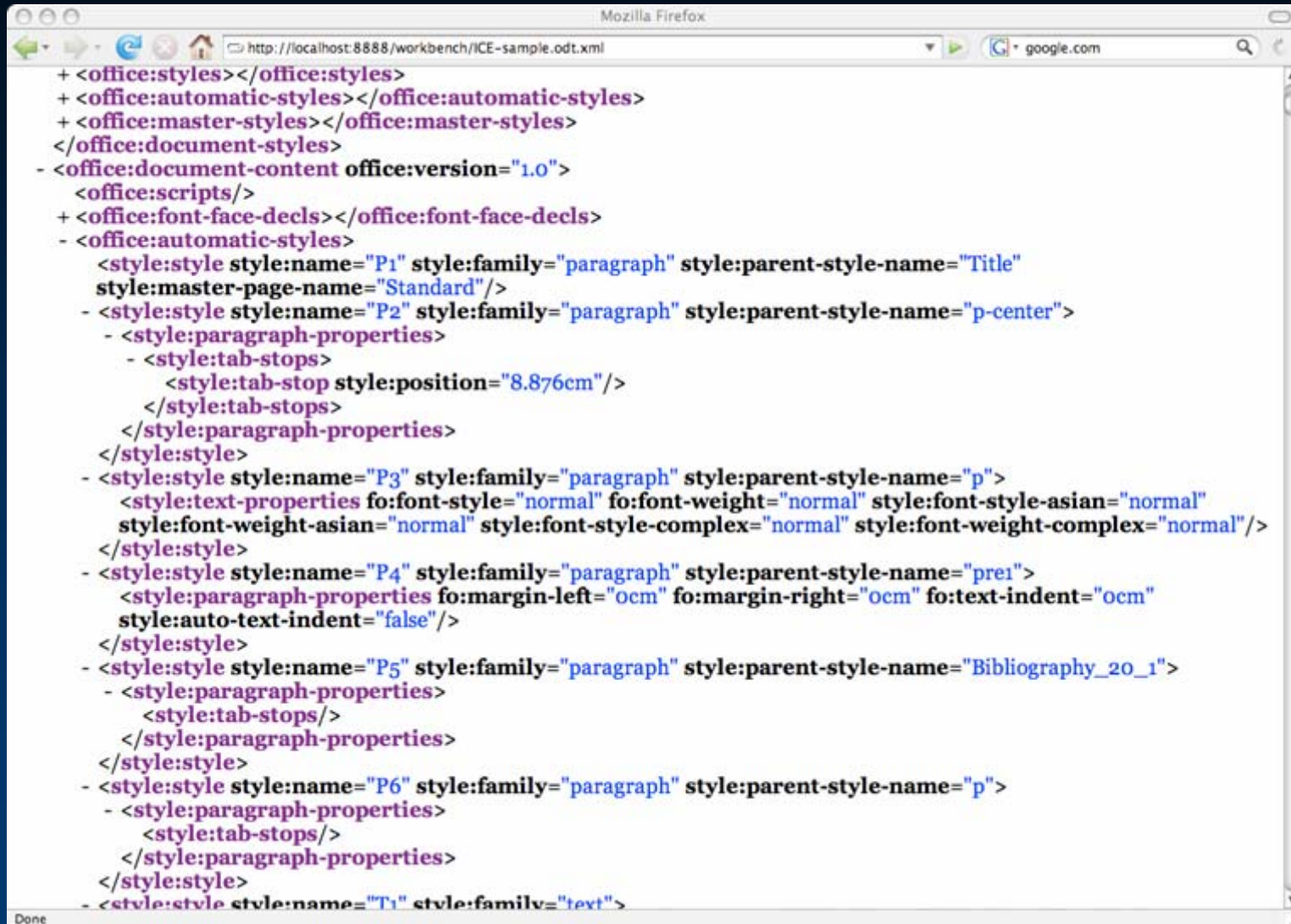
The screenshot shows a web browser window titled "Sample ICE document" with the URL <http://localhost:8888/workbench/ICE-sample.odt.html>. The browser's address bar also shows "google.com". The page content is displayed within a "Digital Scholar's Workbench" interface. The interface includes a navigation sidebar on the left with sections for "Navigation" (Home, APSR, ICE-sample.odt, Pictures, Sauerkraut.jpg, Structure-guesser.odt, Workbench-test-suite), "Document Actions" (Blog this, Version control - register, Version control - revert/sync, Version control - commit, Upload to DSpace, Upload to WebCT, Publish to web site), and "View As" (Plain XHTML, ANU HTML, PDF, DocBook XML, Open Document XML, XSL-FO, Original ODF file). The main content area displays the document's title "Sample ICE document" and author "Ian Barnes" from "The Australian National University". It includes an "Abstract" section with the text "This is the abstract of the document. It should get recognised as metadata and extracted." Below the abstract is a "Table of Contents" listing sections 1 through 12, including "Introduction", "Lists", "Code samples", "Headings", "Images", and "Bibliography". A paragraph of text follows the table of contents, starting with "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?". The document content begins with the heading "1. Introduction".

Open Document Format XML

```
Mozilla Firefox
http://localhost:8888/workbench/ICE-sample.odt.xml
google.com

- <office:document>
- <office:document-meta office:version="1.0">
- <office:meta>
- <meta:generator>
  NeoOffice/2.1$Unix OpenOffice.org_project/680m6$Build-9095
</meta:generator>
<dc:title>Sample ICE document</dc:title>
<meta:initial-creator>Ian Barnes</meta:initial-creator>
<meta:creation-date>2005-09-09T16:03:14</meta:creation-date>
<dc:creator>Ian Barnes</dc:creator>
<dc:date>2007-06-15T09:59:47</dc:date>
<meta:printed-by>Ian Barnes</meta:printed-by>
<meta:print-date>2005-10-21T15:43:43</meta:print-date>
<dc:language>en-US</dc:language>
<meta:editing-cycles>151</meta:editing-cycles>
<meta:editing-duration>P57DT9H9M1S</meta:editing-duration>
<meta:template xlink:type="simple" xlink:actuate="onRequest" xlink:role="template"
xlink:href=" ../.openoffice.org2.0/user/template/ice.stw" xlink:title="ice" meta:date="2005-09-09T16:03:14"/>
<meta:user-defined meta:name="Author">Ian Barnes</meta:user-defined>
<meta:user-defined meta:name="Affiliation">The Australian National University</meta:user-defined>
<meta:user-defined meta:name="Email">Ian.Barnes@anu.edu.au</meta:user-defined>
<meta:user-defined meta:name="Status"/>
<meta:document-statistic meta:table-count="0" meta:image-count="2" meta:object-count="0"
meta:page-count="9" meta:paragraph-count="168" meta:word-count="1758" meta:character-count="9935"/>
</office:meta>
</office:document-meta>
- <office:document-styles office:version="1.0">
+ <office:font-face-decls></office:font-face-decls>
+ <office:styles></office:styles>
+ <office:automatic-styles></office:automatic-styles>
+ <office:master-styles></office:master-styles>
</office:document-styles>
+ <office:document-content office:version="1.0"></office:document-content>
</office:document>
```

Open Document Format XML



```
Mozilla Firefox
http://localhost:8888/workbench/ICE-sample.odt.xml
+ <office:styles></office:styles>
+ <office:automatic-styles></office:automatic-styles>
+ <office:master-styles></office:master-styles>
</office:document-styles>
- <office:document-content office:version="1.0">
  <office:scripts/>
  + <office:font-face-decls></office:font-face-decls>
  - <office:automatic-styles>
    <style:style style:name="P1" style:family="paragraph" style:parent-style-name="Title"
      style:master-page-name="Standard"/>
    - <style:style style:name="P2" style:family="paragraph" style:parent-style-name="p-center">
      - <style:paragraph-properties>
        - <style:tab-stops>
          <style:tab-stop style:position="8.876cm"/>
        </style:tab-stops>
      </style:paragraph-properties>
    </style:style>
    - <style:style style:name="P3" style:family="paragraph" style:parent-style-name="p">
      <style:text-properties fo:font-style="normal" fo:font-weight="normal" style:font-style-asian="normal"
        style:font-weight-asian="normal" style:font-style-complex="normal" style:font-weight-complex="normal"/>
    </style:style>
    - <style:style style:name="P4" style:family="paragraph" style:parent-style-name="pre1">
      <style:paragraph-properties fo:margin-left="0cm" fo:margin-right="0cm" fo:text-indent="0cm"
        style:auto-text-indent="false"/>
    </style:style>
    - <style:style style:name="P5" style:family="paragraph" style:parent-style-name="Bibliography_20_1">
      - <style:paragraph-properties>
        <style:tab-stops/>
      </style:paragraph-properties>
    </style:style>
    - <style:style style:name="P6" style:family="paragraph" style:parent-style-name="p">
      - <style:paragraph-properties>
        <style:tab-stops/>
      </style:paragraph-properties>
    </style:style>
    - <style:style style:name="T1" style:family="text">
```

Open Document Format XML

```
Mozilla Firefox
http://localhost:8888/workbench/ICE-sample.odt.xml
google.com

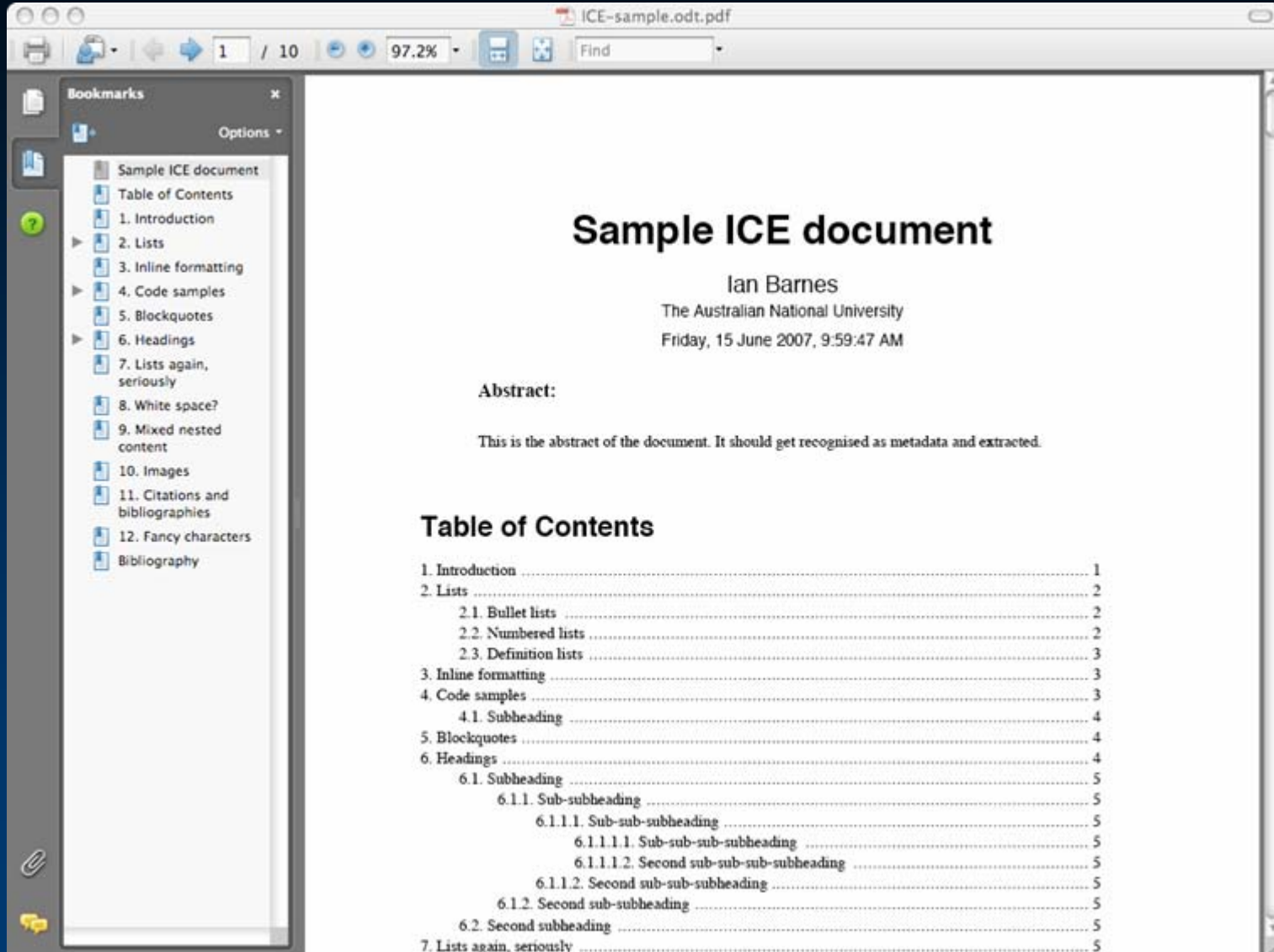
</office:document-styles>
- <office:document-content office:version="1.0">
  <office:scripts/>
  + <office:font-face-decls></office:font-face-decls>
  + <office:automatic-styles></office:automatic-styles>
  - <office:body>
    - <office:text>
      + <office:forms form:automatic-focus="false" form:apply-design-mode="false"></office:forms>
      + <text:sequence-decls></text:sequence-decls>
      - <text:p text:style-name="P1">
        <text:title>Sample ICE document</text:title>
      </text:p>
      - <text:p text:style-name="P2">
        <text:user-defined text:name="Author">Ian Barnes</text:user-defined>
      </text:p>
      - <text:p text:style-name="P2">
        <text:date style:data-style-name="N38" text:date-value="2007-06-15T09:59:46.99">Friday, June 15,
        2007</text:date>
      </text:p>
      - <text:p text:style-name="Abstract">
        <text:span text:style-name="Title">Abstract:</text:span>
        This is the abstract of the document. It should get recognised as metadata and extracted.
      </text:p>
      - <text:p text:style-name="p">
        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?
      </text:p>
      <text:h text:style-name="h1n" text:outline-level="1">Introduction</text:h>
      - <text:p text:style-name="p">
        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
      - <text:note text:id="ftn1" text:note-class="footnote">
        <text:note-citation>1</text:note-citation>
      - <text:note-body>
        - <text:p text:style-name="Footnote">
          I've known him for over 20 years! Our families were in the same ski club when we were kids. And then he went
```

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>
  </footnote>
</sect1>
Done
```

Automatically generated PDF



The screenshot shows a PDF viewer window titled 'ICE-sample.odt.pdf'. The document content is as follows:

Sample ICE document

Ian Barnes
The Australian National University
Friday, 15 June 2007, 9:59:47 AM

Abstract:

This is the abstract of the document. It should get recognised as metadata and extracted.

Table of Contents

1. Introduction	1
2. Lists	2
2.1. Bullet lists	2
2.2. Numbered lists	2
2.3. Definition lists	3
3. Inline formatting	3
4. Code samples	3
4.1. Subheading	4
5. Blockquotes	4
6. Headings	4
6.1. Subheading	5
6.1.1. Sub-subheading	5
6.1.1.1. Sub-sub-subheading	5
6.1.1.1.1. Sub-sub-sub-subheading	5
6.1.1.1.2. Second sub-sub-sub-subheading	5
6.1.1.2. Second sub-sub-subheading	5
6.1.2. Second sub-subheading	5
6.2. Second subheading	5
7. Lists again, seriously	5

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