nature publishing group

Urchin:
A JISC-funded, Open Source Software Development Project

Ben Lund
Nature Publishing Group

OSS Watch
11 December 2003

nature publishing group

Overview

Project Background

- JISC's Request for Proposals
- NPG's Proposal
- RSS and Urchin

Open Source and the development process

- Why Open Source?
- Software design and development

Licences, Copyright, Publication

- Choosing a licence
- Assigning copyright
- Distributing the code

nature publishing group

JISC Request for Proposals

JISC Background

- "help to provide a range of services, tools and mechanisms for colleges and universities to exploit fully the value of online resources and services."

Metadata and Interoperability

- "help to develop interoperability between publishers and aggregators"
- "Proposals are invited to undertake projects working to release metadata about their content"

nature publishing group

NPG's Proposal

Focused on RSS (RDF Site Summary)

- Increasingly popular means of information dissemination on Web
- Has great metadata carrying potential

Identified three barriers to adoption

- Information providers must write custom code to publish RSS feeds
- Hard for non-programmers to merge and filter RSS feeds
- Harder still for publishers to set up RSS based news aggregation services for particular areas of interest

Develop an Open Source, Web-based RSS aggregator and filter.

nature publishing group

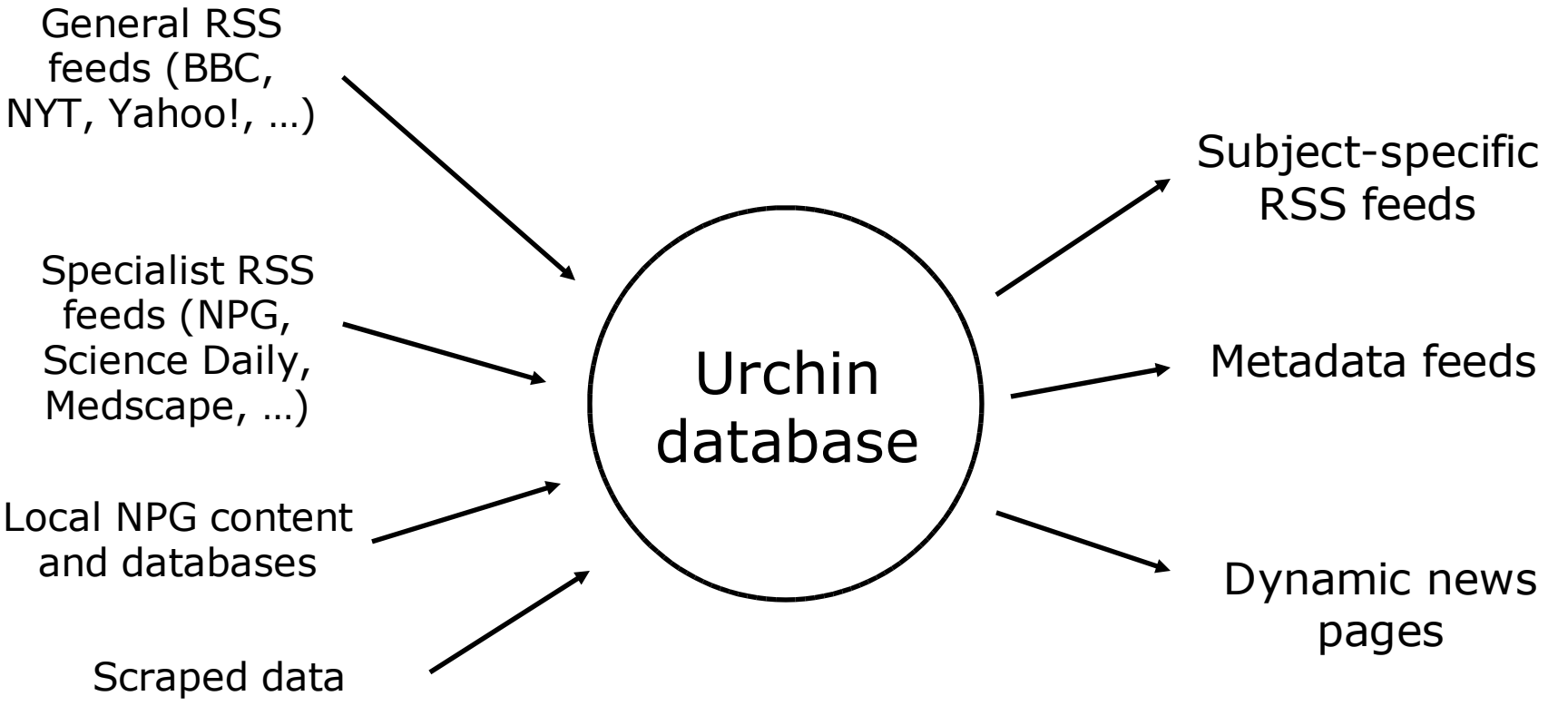
What is RSS?

- XML format for lightweight article syndication
- Core information:
  - List of current articles
  - Titles, links, descriptions
  - May optionally contain metadata about the articles and the Website
- XML file is published on the Web

```
<?xml version="1.0" encoding="utf-8"?>
<?xml-stylesheet href="/styles/rss.css" type="text/css"?>

<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns="http://purl.org/rss/1.0/">

<channel rdf:about="http://www.nature.com/nsu/rss.rdf">
<title>Nature Science Update</title>
<description>The latest science research and news reported by Nature's science writing
team</description>
<link>http://www.nature.com/nsu/</link>
<image rdf:resource="http://www.nature.com/nsu/slices/rss_logo.gif"/>
...
</channel>
...
<item rdf:about="http://www.nature.com/nsu/031124/031124-14.html">
<title>United Nations climate meeting begins in Milan</title>
<link>http://www.nature.com/nsu/031124/031124-14.html</link>
<description>Clean development and carbon sinks set to dominate COP9 policy agenda.</description>
<dc:date>2003-12-01</dc:date>
<dc:creator>John Whitfield</dc:creator>
</item>

<item rdf:about="http://www.nature.com/nsu/031124/031124-10.html">
<title>Greenhouse gases level off</title>
<link>http://www.nature.com/nsu/031124/031124-10.html</link>
<description>Concerted efforts could further fall of methane.</description>
<dc:date>2003-11-28</dc:date>
<dc:creator>Helen Pearson</dc:creator>
<dcterms:references>doi:10.1029/2003GL018126</dcterms:references>
</item>
...
</rdf:RDF>
```

nature publishing group

Urchin Design



Multiple input formats and sources → Multiple output formats and content subjects

nature publishing group

Why Open Source?

Aims of project required it

- help to overcome barriers to adoption of RSS
=> broadest possible user base
- individual modules independently useful
=> full access to code

Building on other Open Source tools

- Using much pre-existing Open Source code (XML::RSS, RDF::Core, XML::XSLT, Parse::RecDescent)
- Licences don't necessarily mandate Open Source, but…
- May need to modify them

Good for NPG

- The more Scientific RSS feeds, the better services we can provide

Publicly funded work

nature publishing group

Open Source and Design (1)

- Wanted to reduce amount of code writing required for smaller publishers to issue RSS feeds
  - Individual Urchin modules independently useful
  - Standard interfaces to those modules
- Other implementers would have unforeseen requirements
  - Highly customisable output styles
  - Pluggable query syntax grammar
  - Database schema to cover all possible input data (!)

nature publishing group

Open Source and Design (2)

- Open Source web application => Run on (at least) an open source platform

  Developed using:

  - Perl
  - MySQL

  Developed on:

  - Apache 2
  - Red Hat Linux 8.0

- Open Source affected the design (but in a good way…)

- Open source affected the platform choice (but we would have used those tools anyway…)

nature publishing group

Open Source and Development

- Pre-existing Open Source code greatly simplified development
  - Very powerful search syntax developed in a few hours using Parse::RecDescent
  - Specialised RSS and RDF tools available
- Limitations of that code restricted initial functionality
  - RSS output limited by what XML::RSS can produce
  - Good news is we can patch it in the future…

nature publishing group

Licence Choices (1)

GNU General Public Licence (GPL)

- Can copy and redistribute
- Can modify and redistribute
- No Warranty
- Preserves original copyright and licence terms

GNU Lesser General Public Licence (LGPL)

- As GPL
- Allows code to be used as library in non-Open Source application

Artistic Licence

- As GPL, except:
- Non-Free / Open Source derivatives must be distributed under a different name to original copyright holder's version

nature publishing group

Licence Choices (2)

- We chose LGPL for independent modules
  - Modules designed to lower barrier to producing RSS
  - Designed to be used outside Urchin
  - LGPL increases potential usage
- We chose GPL for application code
  - Query interface and admin tools
  - Keep it Free
  - Maximum benefit from others work

nature publishing group

Copyright

- "Any information gathered during the course of the demonstrator and not already in the public domain is deemed to be the property of JISC."
- External programmer wrote code, paid for by JISC

Copyright of the code assigned according to JISC's wishes:

Copyright (C) Higher Education Funding Council for England (HEFCE), 2003

nature publishing group

Distributing the code (1)

- Central Repository or Institutional website?
  - http://sourceforge.net - **72,393** hosted projects
  - http://savannah.nongnu.org/
- Central repository offers
  - CVS code hosting
  - Distribution files and documentation hosting
  - Download statistics
  - **Project visibility**

nature publishing group

Distributing the code (2)

Application for space on SourceForge

- Submit short, public description
- Submit detailed description of proposed project
- State Licence(s) used (must be OSI-approved)
- Names of languages, libraries, databases, frameworks used
- Names of platforms supported
- Purpose of software
- Major features
- Discuss major challenges faced in development
- Information about existing work

nature publishing group

Documentation

Complete documentation required by our agreement with JISC

- Installation and usage
- Customisation instructions

Better than some Open Source projects…

- "the documentation is good, especially the installation guide"
- "It is a common complaint against many Open Source projects that the code is both obscure and poorly documented. This is not the case with the Urchin modules."

nature publishing group

Further Work

- JISC commissioned independent report on software
  - Highly complimentary
  - Recommended further funding
- Both NPG and JISC are funding further work
  - New code will be Open Source
  - Will use SourceForge CVS
  - Focus is still Metadata and Interoperability

nature publishing group

More Information

PALS Metadata and Interoperability Projects

http://www.jisc.ac.uk/index.cfm?name=programme_pals

Urchin Website

http://urchin.sourceforge.net/

Email

b.lund@nature.com