

# Intute Repository Search Project

## Overview

**Vic Lyte** FHEA  
IRS Development Manager

**June 2007**

## **Pilot phase - Technical review & evaluation**

- Existing aggregator services provide similar simple search functionality
  - little in the way of interesting value-added services;
  - Don't want to try to "out-Google" Google
- Development of a Demonstrator based on ePrints (<http://irs.ukoln.ac.uk>);
- Need to scale to service-level and increased functionality;
- Major issue will be how to automatically access full text documents for value-added services;
- Need for richer metadata application profile for OAI export;
- Need for semantic (contextual) exploration.

## Potential development paths from evaluation

- *Promote a more complex metadata format for OAI export and develop 'plug-ins' for downloading;*
- *Full-text indexing of documents;*
- *Personalisation;*
- *Development of embedding services;*
- *Experiment with text-mining full-text documents;*
- *Consider approaches to automatic subject classification;*
- *Investigate name authority issues;*
- *Support for Web 2.0 services based on aggregated and dynamic taxonomy content?*

## Vision

- *Expose the outputs of research, learning and teaching so that it is visible and usable to the benefit of future UK research / Learning & Teaching Communities.*

## Mission

- *To link and expose repositories by the exploitation (and exploration) of available search / technologies and structured metadata approaches.*
  - *Achieved over a 3-year period through three negotiated and planned iterations focussing on desired end-user and stakeholder benefits.*

## Scope

- *The scope of the IRS Project is intended to initially cover both a UK national and where appropriate, global dimension and support the following domains:*
  - **Research Lifecycle** (discovery, development, collaboration, dissemination);
  - **Teaching & Learning** (resources, pedagogic activity / processes, resource-based learning);
  - **Research Administration** (deposition, repository support and exposure)

Over a 3-yr period to identify and develop:

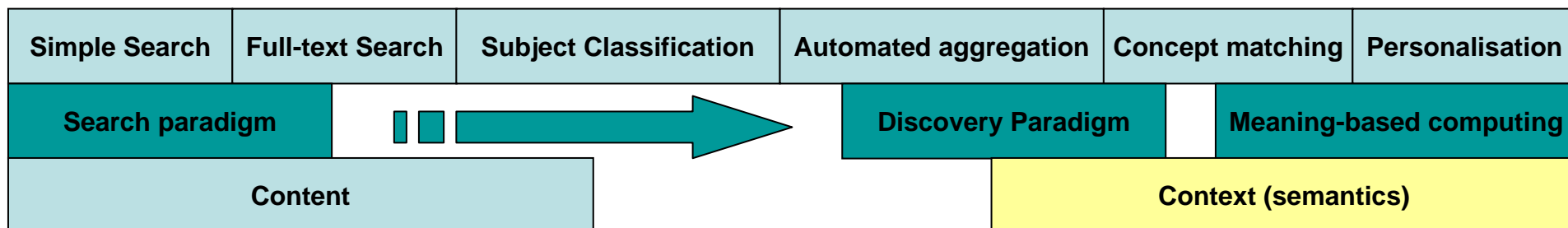
- **Cross repository search**, aggregation and retrieval from all HE and relevant UK repositories;
- Development of **machine to machine** interface
- Exploration and resolution of issues relating to achieving **full-text searching**, discovery;
- Achievable synergies between research and **learning object repositories**;
- Opportunities from **international collaboration** in this area (UK / EU);
- **Scalable and flexible** search infrastructure / service supporting a number of stakeholder constituencies;
- How IRS can support agreed value-added **personalisation features** (JISC);
- How the service can support / **compliment allied programmes** in obtaining cultural acceptance and embedding into day-to-day Research Desktop environment;
- Tools to establish **metrics** for cross-searching / support for research appraisal process;
- **'Showcase'** for collective and collaborative UK research output.

## Knowledge Management Context for Researchers, Teachers and Students



### Knowledge Context

- Where can I find...?
- What can help me?
- Who can help me?
- What do we know?
- What do I / we don't know?

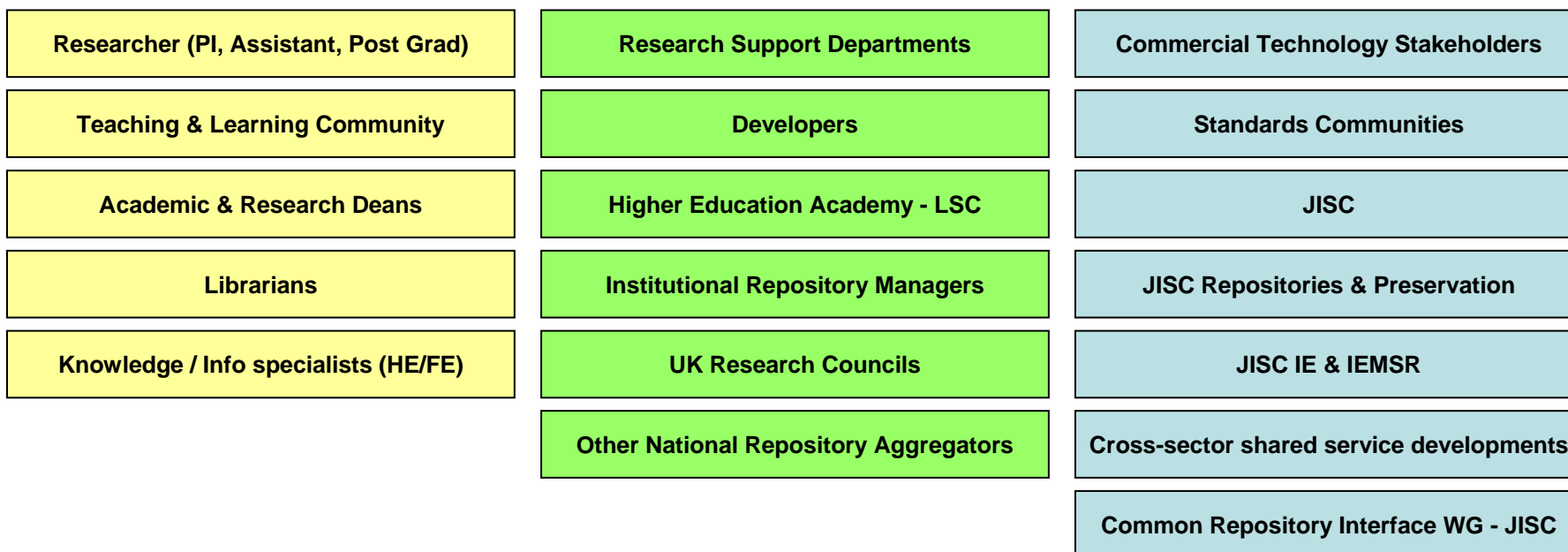


## **Phase #1 will focus on:**

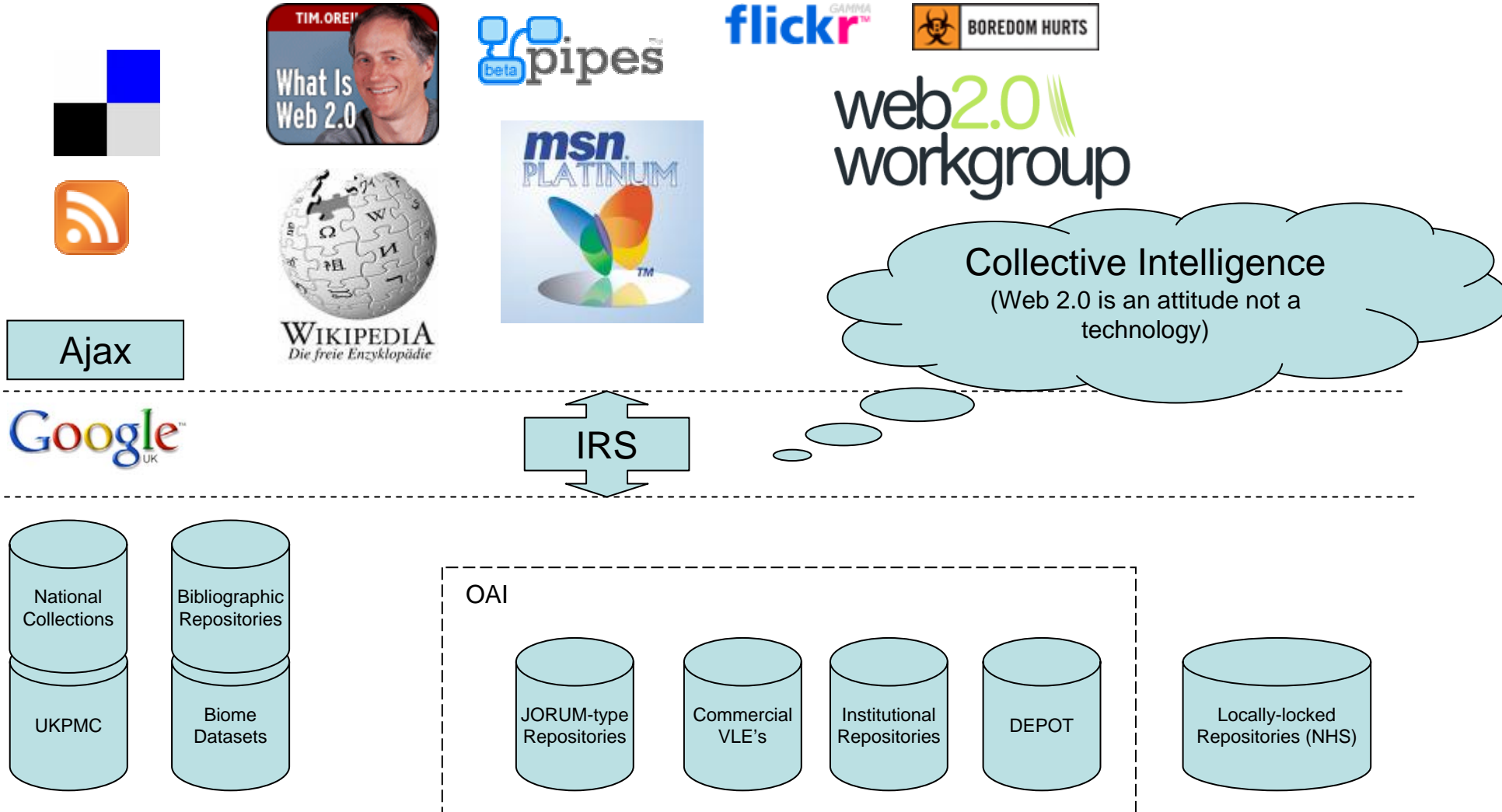
- *Extend demonstrator (61 UK Academic Repositories, 150,000 Working Papers etc;*
- *Capture, analysis and management of scenario-based requirements;*
- *Technology Watch*
- *Modelling and design activity;*
  - *Service in development;*
  - *Critical use cases.*
- *Scale and deploy simple search functionality / harvesting across the range of UK repositories;*
- *Identify approach to subject searching and Learning Object Repositories;*
- *New web interface;*
- *Upgrading technical architecture (Lucene / ?Harvester)*

**Phase #1 will focus on:**

- *Capture, analysis and management of scenario-based requirements (stakeholders and end-users);*



- Rapidly changing search environment

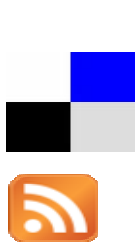


**Phase #2 will focus on:**

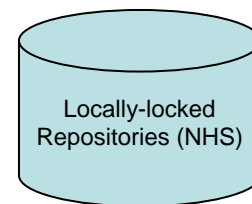
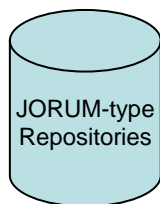
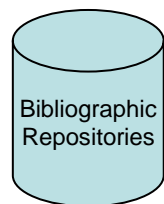
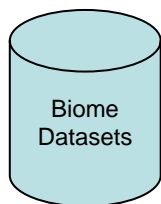
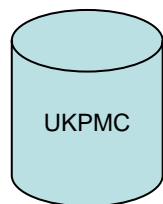
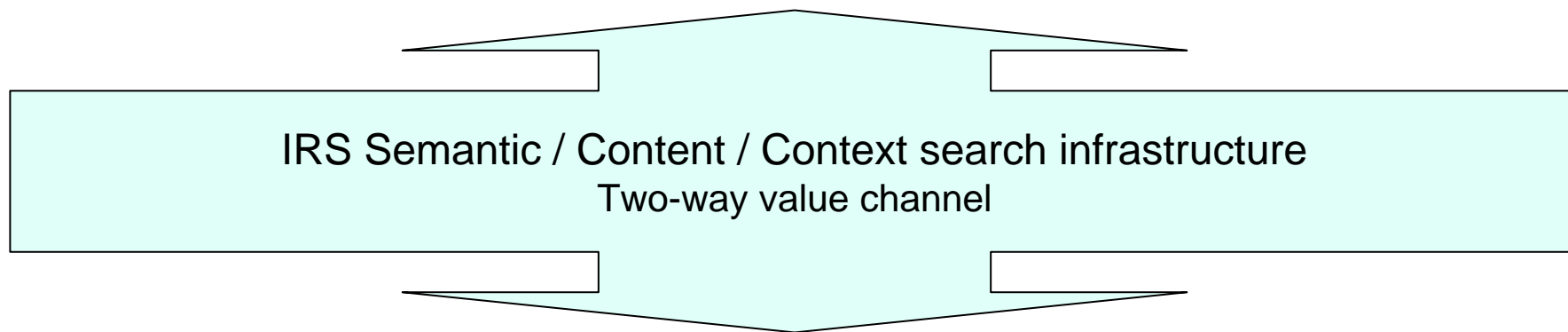
- *Agreement and prioritisation of requirements for relating to critical use cases;*
- *Elaboration and construction activities relating to meeting achievable requirements for extended functionality such as:*
  - *personalisation and embedding features (i.e. SOAP).*
- *This will be constrained by dependencies on related technologies, projects and initiatives.*

**Phase #3 will focus on:**

- *Elaboration of scenario-based requirements related to areas such as discovery, text-mining, aggregation and profiling;*
- *Establishing a feature mapping between all technologies available to the Project at that time against requirements derived from agreed critical use cases;*
- *Development of a Gap Analysis if appropriate;*
- *Fully-costed Options Appraisal and recommendations for next-stage development priorities.*



## Portals

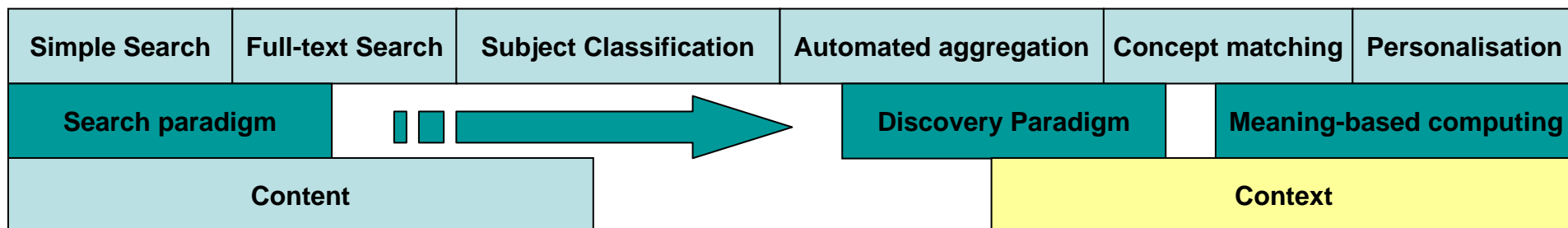


## Knowledge Management Context for Researchers, Teachers and Students



### Knowledge Context

- Where can I find...?
- What can help me?
- Who can help me?
- What do we know?
- What do I / we don't know?





## Suggested areas:

- *Subject searching - requirements gathering and exploration of feasible approaches at appropriate stage in the deposit to discovery lifecycle;*
- *Links with international projects/initiatives - joint information gathering and setting up strategic alliances;*
- *Advocacy - including sharing plans and approaches and findings, also joint events, conference papers, publicity materials i.e. very practical efficiency savings;*
- *Standards - development and advocacy;*
- *Repository landscape - making sense of it together in order to prioritize strategically and identify quick wins (e.g. prioritizing search targets);*
- *Lobbying repository search software suppliers for any changes to enhance the output of our projects e.g. adopting standards;*
- *Sharing links and experiences of related work e.g. UK PubMed, DRIVER and sharing experience outputs relevant to other projects in the programme.*

## Suggested Phase III development areas:

- **enhancing and augmentation of metadata creation** via automatic (machine-driven) classification / meaning-based taxonomy extraction; and particularly adding value to metadata through extraction of key facts from text, represented as instances of relations between concepts;
- **automated classification and clustering;**
- **cross-disciplinary (subject) metadata extraction** to expose common interdisciplinary areas which would be of high value in teaching and learning contexts;
- development of algorithms that perform query expansion by **grouping semantically similar concepts** which can be used in searching across different disciplines;
- development of **algorithms that disambiguate concepts between disciplines**, for example the term stress in nursing and in material science, denotes different concepts but share the same form;
- development of **aggregated concepts to support visualization requirements** such as tagged cloud views (e.g., as in [www.quintura.com](http://www.quintura.com)). This will be of high value to provide support to social and qualitative modes of inquiry to offer parity with investments being made to support **quantitative** and life science-oriented research areas;
- **summarization;**
- supporting institutions in their classification efforts through **development of an auto classification tool**. This would support the work of cataloguers by offering feedback to enhance their service provision and role as knowledge-support personnel.