Pistoia Alliance

Emerging Life Sciences Collaboration on Common Service Specification

Ian Harrow (Pfizer) and Nick Lynch (Astra Zeneca) for the Pistoia Alliance
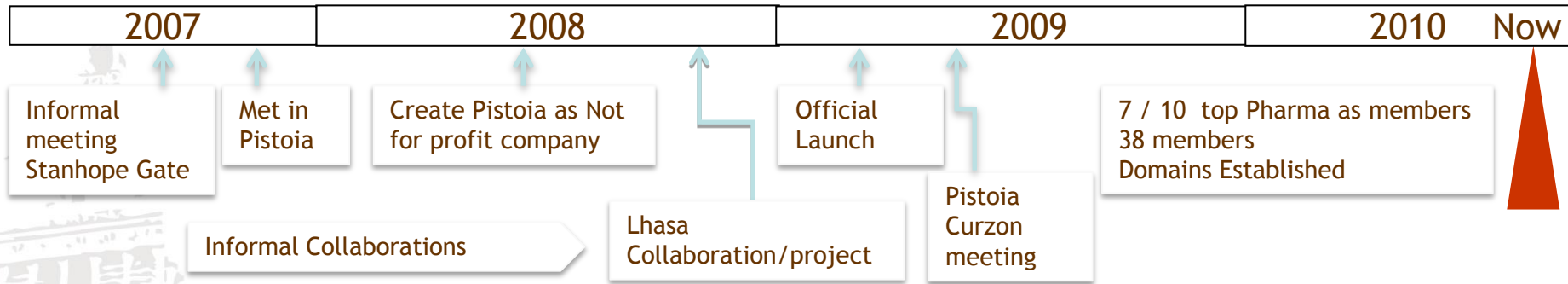
ICIC 2010

26th Oct 2010

http://pistoiaalliance.org

Pistoia Alliance

Presentation Outline

- Pistoia Organisation
- Four projects:-
  - Biomedical Knowledge Brokering SESL pilot
    - More depth on this
  - Vocabulary Standards Initiative
    - An emerging project
  - Sequence services
  - Electronic Lab Notebook
- Summary
- Acknowledgements

Pistoia Background and History

Timeline: 2007 — 2008 — 2009 — 2010 — Now

- Informal meeting Stanhope Gate
- Met in Pistoia
- Create Pistoia as Not for profit company
- Official Launch
- 7 / 10 top Pharma as members, 38 members, Domains Established
- Informal Collaborations
- Lhasa Collaboration/project
- Pistoia Curzon meeting

Pistoia Description

The primary purpose of the Pistoia Alliance is to streamline non-competitive elements of the life science workflow by the specification of common standards, business terms, relationships and processes

Pistoia Goals

- to allow this framework to encompass/support most pre-competitive work between the organisations
- to support life science workflow prior to submission
- to work with other Standards organisations

History

Initial Meeting with GSK, AZ, Pfizer and Novartis – outlined similar challenges and frustrations in the Informatics sector of Discovery

The advent of Web Services and Web 2.0 allow for decoupling of proprietary data from technology

Publicly available structural and biological DBs allow for a non-IP related analysis and as a scientific test suite.

Sponsorship from R&D IS heads within Life Science industry
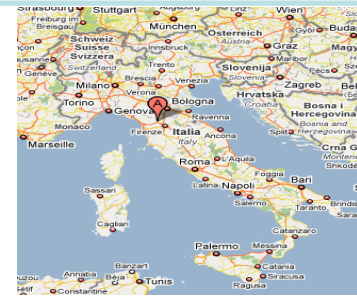
Pistoia Domains

Pistoia Domains group areas of interest, scope and deliver projects

Pistoia Groups – as defined in byelaws

- Board of Directors
- Officers (Operational Team)
- Technical Committee

Pistoia Domain – high level collection of Working Groups with common themes

Domain Steering Groups

Allows governance across a domain using Working Group chairs and Technical Committee reps

Working Groups

The main project delivery mechanism in Pistoia. All standards will be delivered by WGs

Pistoia Members

Provide expertise for WGs and running Pistoia
Define:
- Requirements
- Technical Standards
- Service Standards

External Groups outside of Pistoia

Could:
- Join Pistoia
- Influence Pistoia members
- Influence through other standards groups and activities
- Collaborate on standards' feasibility studies
- Collaborate through non-Pistoia Standards initiatives

Pistoia Alliance

Pistoia Membership

Pistoia Domains

Pistoia Domains focus on business workflows /supply chains

Enabling

- Vocabulary
- Visualisation
- Workflow
- Others

Knowledge and Information Services

VSI   SESL

Application Integration

Biology Data Services	Chemistry Data Services	Translational Data Services

Sequencing   ELN
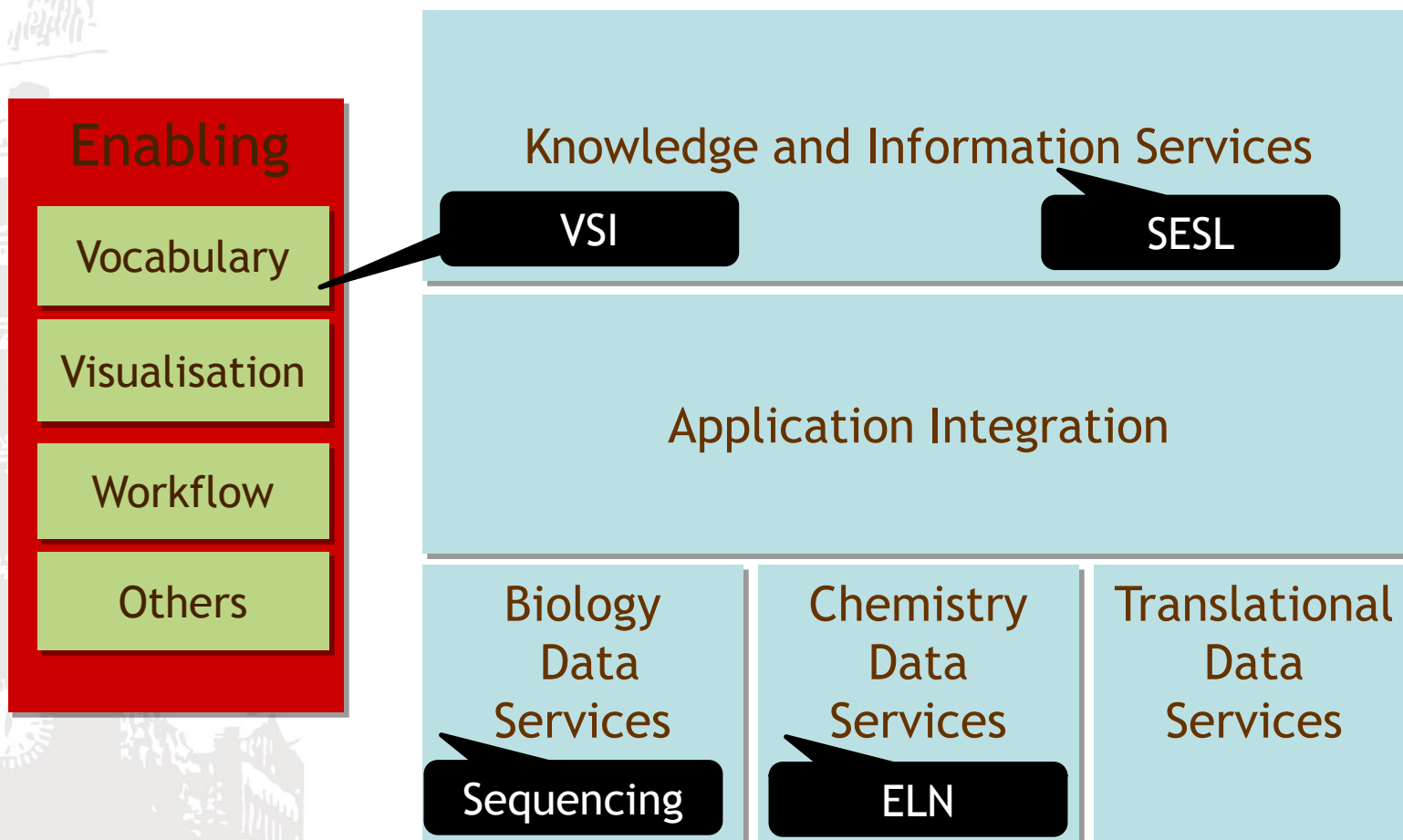
The Pistoia SESL Project

Pistoia Alliance SESL Pilot for a Biomedical Brokering Service

Ian Harrow, Wendy Filsell, Dietrich Rebholz Schuhmann

http://pistoiaalliance.org

Pistoia Alliance

SESL: Biomedical Knowledge Brokering

- **Challenge:**
  - No single system for retrieving gene to disease relationships contained in both published & biological database content
  - Need a 'push model' for biomedical knowledge access: the current model requires the consumer to search 1000's of content sources
- **Opportunity: Pilot Project with key stakeholders**
  - Pilot a 'push model' for biomedical knowledge brokering
  - Engage multiple consumers, content providers and a single, public group to develop the necessary infrastructure to explore the standards required for the model to work in production
- **History:**
  - May 2008: Common Disease Knowledge Environment (CDKE) IMI call drafted
  - Sep 2008: postponed call publication
  - Jan 2009: x-pharma meeting in London on how to progress CDKE
  - Apr 2009: CDKE presented at SESL workshop
  - Oct 2009: SESL Pilot meeting (funders)
  - Jan 2010: Pilot launch

The Knowledge Service Framework

Multiple Consumers

Knowledge Applications

Common Service Broker

Content Suppliers

'Consumer' Firewall

Open Stds

Supplier Firewall

Disease Dossier

Service Layer
Assertion & Meta Data Mgmt
Transform / Translate
Integrator

Std Public Vocabularies

Business Rules

Corpus 1

Db 2

Db 3

Db 4

Corpus 5

Effort required to fit DBs to service layer
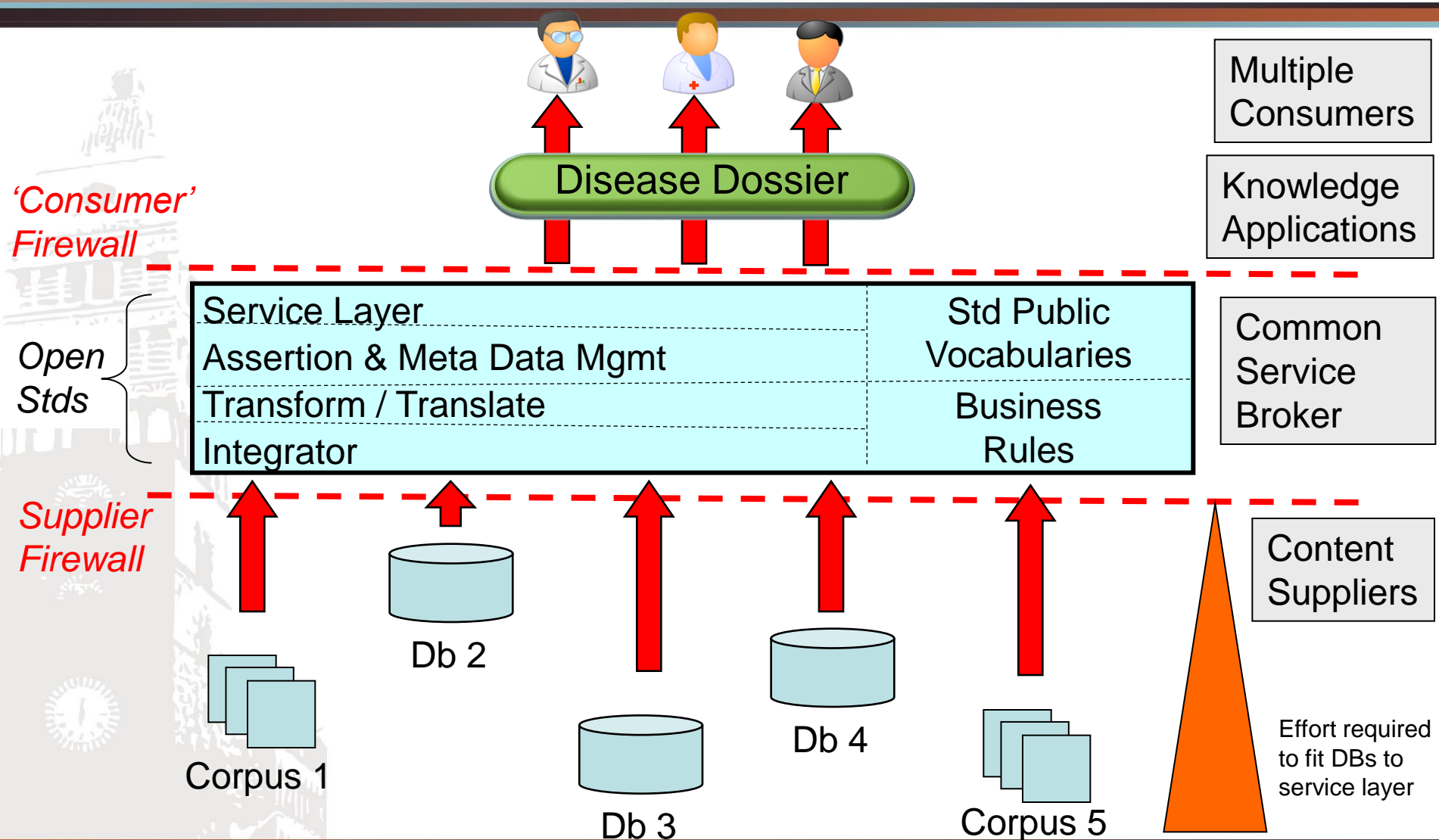
A Production Service vision…
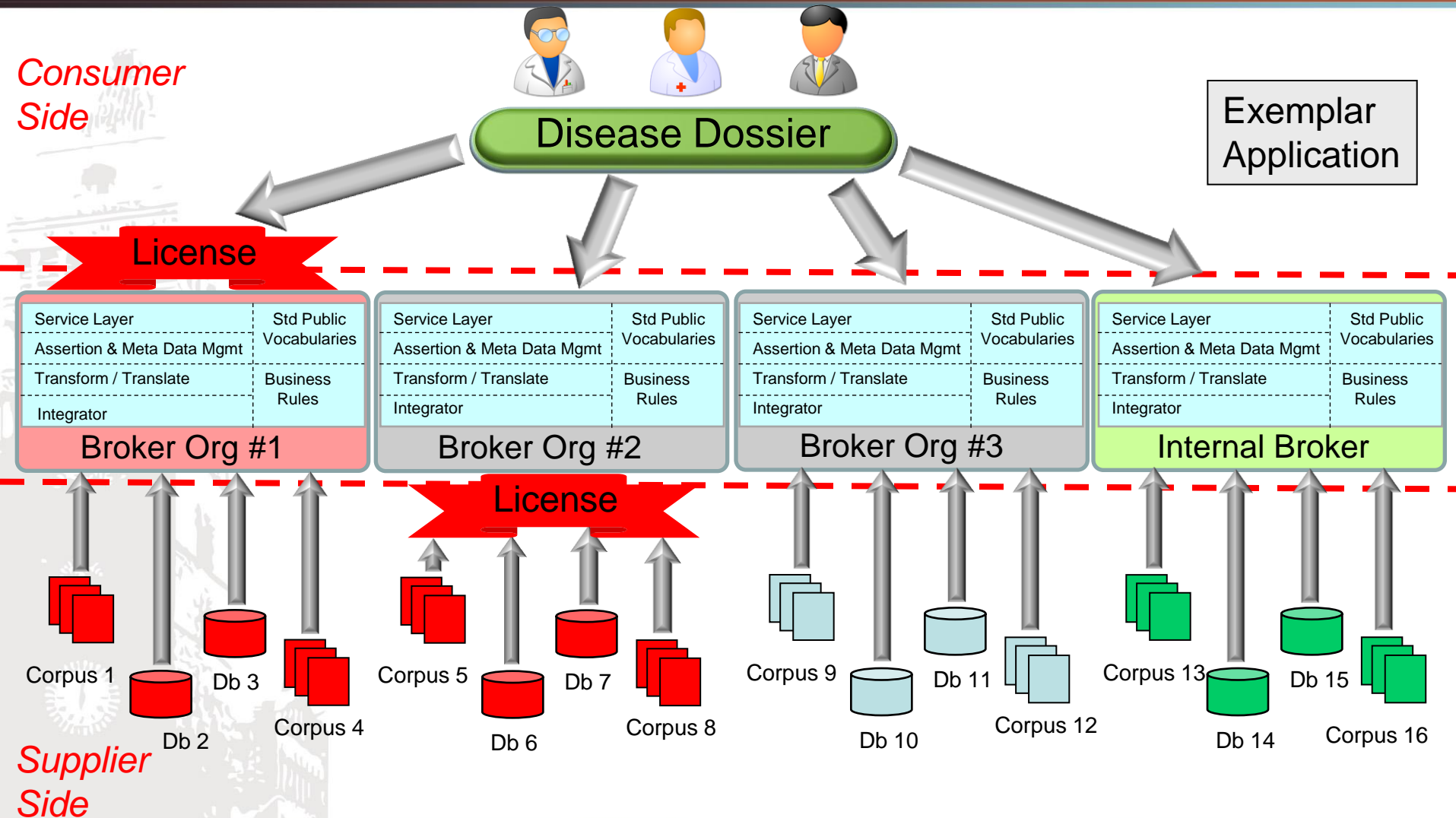
The Pilot

- **Deliverables:**
  - Publication of standards & recommendations for service implementation
  - Pilot implementation of service for a single disease (assertions from pre-defined document sets & databases)
  - **Establish ways of working pre-competitively across industry/vendor/academia**
  - Dialogue and assessment of cost / value, with key content suppliers in moving to such a push model for content (viability of moving to production)
- **Status:**
  - **AZ, Pfizer, GSK, Roche, Unilever, EBI, NPG, OUP, Elsevier & RSC**
  - 12 month project, £200K direct funding (+ PM & Architecture support)
  - Contract between Pistoia & EBI signed 20th January 2010 for 1 year
- **Scope:**
  - Development of an assertion database in combination with a user interface and associated web services for one disease/indication/phenotype of broad interest: **Type II Diabetes**
  - Assertional content derived from 3 structured data sources and limited Journal content (co-occurrence & statistical derivation from full text)
  - Assertional evidence for filtering and drill down to primary data.
  - Limited vocabulary development for area of focus: **Type II Diabetes**

Minimal configuration to test a Brokering Service

User Interface

Interface Layer at consumer org'n

Condition:
Identical structure.
Different content
which can overlap.

Broker #1

Service Layer	Std Public Vocabularies
Assertion & Meta Data Mgmt	
Transform / Translate	Query templates
Triple store 1	

Broker #2

Service Layer	Std Public Vocabularies
Assertion & Meta Data Mgmt	
Transform / Translate	Query templates
Triple store 2	

Brokering service Layer at EBI

Primary source Layer at provider org'n

EBI Swissprot database

NCBI OMIM

Elsevier corpus

RSC corpus

EBI Array Express database

EBI Swissprot database

NPG corpus

OUP corpus

Pistoia Alliance

SESL user interface mock-up

SESL Results

	Gene	R'ship	Disease	Species	Evidence
1	abc1	Co-occurs	Diabetes	Mus	*Paper UID:1234*
2	abc1	Up-Reg	Diabetes	Homo	*ArrayExpress: XXX*
3	abc2	Co-occurs	Diabetes	Homo	*Paper UID:1344*
4	abc13	Co-occurs	Diabetes	Mus	*Paper UID:1314*
5	abc7	Mutation	Diabetes	Rattus	*OMIMI: XXX*
6	abc1	Co-occurs	Diabetes	Mus	*Paper UID:45643*
7	abc1	Co-occurs	Diabetes	Homo	*Paper UID:2143*
8	abc1	Co-occurs	Diabetes	Mus	*Paper UID:1204*

Pistoia
Alliance

Timelines: Development Phase

Pistoia Alliance

Task/Deliverable	Phase	Type	Jan-10	Feb-10	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10	Jan-11	Feb-11
			Month 0	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12	
Finalised Technical Specification document (Month 4)		**Deliverable 1**					^ ✓									
Build vocabularies within scope	Development	Task 2								✓						
RDF data export from UniProt and Ensembl	Development	Task 3								✓						
RDF data export of Array Express	Development	Task 4								✓						
Extract literature assertions for T2DB from publishers' content	Development	Task 6									✓					
Develop RDF triple store schema and demonstrator	Development	Task 7										✓				
Develop query definitions	Development	Task 8										✓				
Establish API services for remote access	Development	Task 9										✓				
Develop simple user interface for demonstrator (based on mock-up)	Development	Task 10										✓				
Write documentation that defines the standard framework	Development	Task 11										✓				
Access to early prototype demonstrator and report (Month 7 & 8)		**Deliverable 2 & 3**									^^ ✓					
Final prototype demonstrator, recommendations post-pilot, report (Month 11 & 12) and public launch		**Deliverable 4 & 5**												^	^ ^	

Timelines:
Testing and Communication Phase

Task/Deliverable	Phase	Type	Jan-10	Feb-10	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10	Jan-11	Feb-11
			Month 0	Month 1	Month 2	Month 3	Month 4	Month 5	Month 6	Month 7	Month 8	Month 9	Month 10	Month 11	Month 12	
Tests of the demonstrator (full private and limited public instance)	Testing and communication	Task 12									▓	✓	▓			
Deploy publc demonstrator	Testing and communication	Task 13										▓	▓	▓		
Write publication for standard definition	Testing and communication	Task 14											▓	▓	▓	
Develop recommendations for post-pilot project	Testing and communication	Task 15												▓	▓	
Final prototype demonstrator, recommendations post-pilot and report (Month 11 & 12)		Deliverable 4 & 5												^	^	
Public release of limited demonstrator (Month 13)		Deliverable 6														^

Summary for SESL pilot

- Significant progress to towards realising the technical goal of knowledge brokering
  - Can a push model work? A **"**hyper**"**standard?
- A unique consortium from three cultures: industry, publishers and academia
  - Working together – <u>sharing</u> costs and risks
- Business opportunities and concerns
  - For data providers and consumers?
- Phase 2 planning is underway for 2011

The Pistoia VSI Project

Vocabulary Standards Initiative

Project Leads: Lee Harland and Christopher Larminie

http://pistoiaalliance.org

Pistoia
Alliance

Standardizing Drug Target Types

- Representation of a molecular drug target in structured databases is ad-hoc
  - Single protein-targets are "OK" (being linked via Entrez gene, but this is not an agreed standard)
  - Multi-protein targets, complexes, biologicals and many more are poorly described, often simply raw text
- This project will focus on industry & suppliers to describe a specification for reporting drug targets within structured content
  - Minimal cost, just FTE time required
  - This could feed into the IMI Open Pharmacology (OPS) call as an industry-publisher requirement
  - Output would be a specific set of "rules" regarding the representation of complex molecular targets
  - Aim would not be to define a list of all known targets, this would be out of scope. As will any text-mining efforts.
  - Recommendation to suppliers and industry to adopt specification along with industry-generated mappings for pre-existing targets
  - Deliverable – specification & publication
- Could be a start to a future, wider pharmacological data standard project
  - All databases providing pharmacological activity content delivered in a standard way
  - Could gain a quick-start building on MIABE standard

The Pistoia Sequence Services Project

Project Lead : Simon Thornber

http://pistoiaalliance.org

Pistoia
Alliance

Sequence services Project

Description

As a drive to cuts costs, encourage standards, and provide simplification it is proposed that Pistoia commission a set of secure internet hosted sequence services.
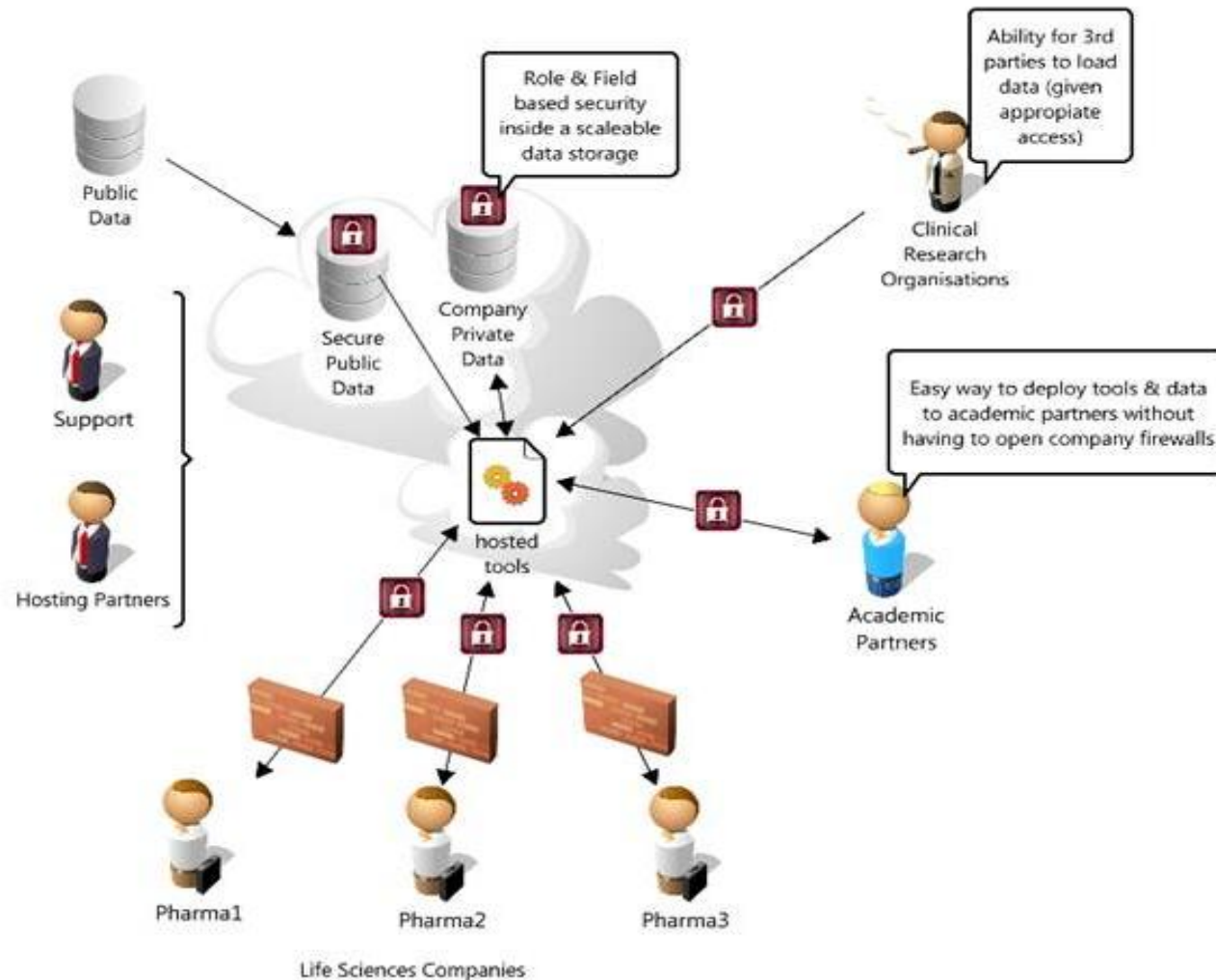
Benefits

These services will ultimately provide access to public, private & commercial data & tools, that will enable scientists to search, store & analyse all their sequence based data in a single web interface.

Current Status for sequence services

- Defined the Project Vision
- Split Vision into achievable phases of delivery
- Defined Phase 1 use cases
- Focus on Non-Functional use cases e.g. security
- Scoring criteria in final stages of drafting
- 5 Vendor presentations during May / June 2010
  - Cognizant +Eagle Genomics, ThomsonReuters, Genome Quest, & Constellation Technologies + Microsoft + AWF and the STFC.

The Pistoia ELN Project

Project Lead : Richard Bolton

http://pistoiaalliance.org

Pistoia
Alliance

ELN Project Description and Benefits

Description

To deliver a query service standard applicable for use with data types commonly found in electronic lab notebooks (ELN's). The initial scope will be against chemistry related ELN's but the solution should aim to be general enough that it can be applied to other scientific notebook applications.

Benefits

Searching of data stored in ELN's from different vendors. Lowering the costs of using ELN data with partners and CRO's.

Current Status for ELN

- Active Participation at biweekly meetings from GSK/AZ/Pfizer/BMS/Symyx/Edge/Accelrys
- Agreed 3 delivery phases
- Phase 1 Definition of problem space and creation of users stories.
  - Complete. User Story Document 'published'
- Phase 2 Creation of ELN Query services definition.
  - End to end process run through by team to create a full model for two of the user stories.
  - GGA chosen to complete work. Funding agreed and approved by operations team. Work started but contract not yet in place.
- Phase 3 Creation of POC in partnership with Vendor.
  - Not yet started. Will likely require vendor partnership, budget and technology decision.

ELN Summary
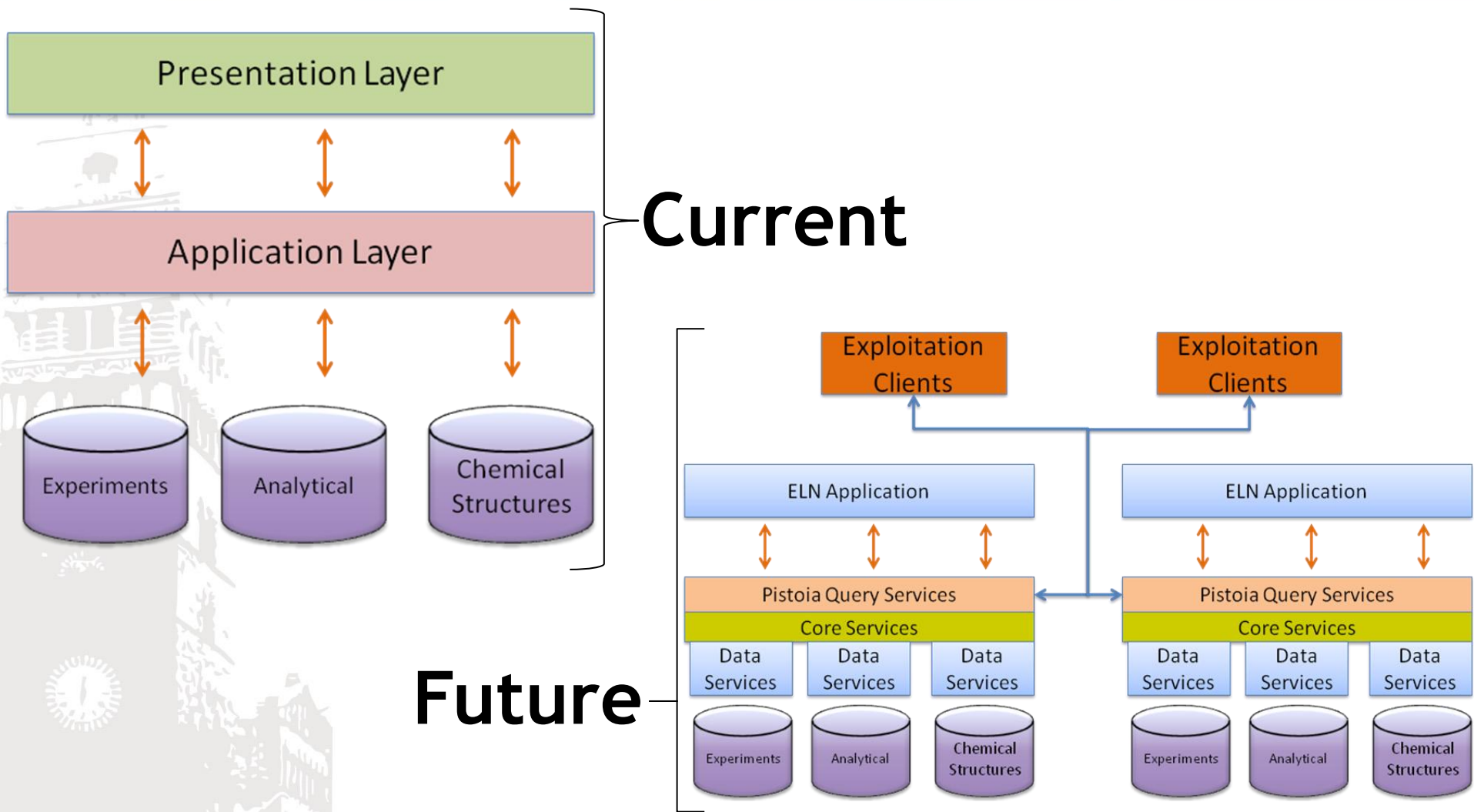
Summary for Pistoia projects

- ## SESL Biomedical Knowledge Brokering
  - Phase 1 pilot to complete by end 2010
  - Phase 2 is planned
- ## Vocabulary Standards Initiative
  - An emerging project on Drug Targets
- ## Sequence services
  - Phase 1 nearing completion and Phase 2 planned
- ## Electronic Lab Notebook
  - Phase 1 is complete and Phase 2 is underway

Acknowledgements

SESL

Dietrich Rebholz Schumann, EBI
Silvestras Kavaliauskas, EBI
Christoph Grabmuellerm EBI
Dominic Clark, EBI
Mike Westaway, AZ
Ian Dix, AZ
Wendy Filsell, Unilever
Ian Stott, Unilever
Peter Woollard, GSK
Nigel Wilkinson, Pfizer
Catherine Marshall, Pfizer
Michael Braxenthaler, Pfizer
Jabe Wilson, Elsevier
Richard O'Bierne, Oxford UP
Richard Kidd, RSC
Alf Eaton, Nature PG

ELN

Richard Bolton, GSK
David Drake, AZ
Steve Trudel, Pfizer
John Duncan, Pfizer
Uwe Geissler, Novartis
Carol McNab, BMS

Vendor reps from:-
Symyx
Edge
Accelrys

Sequencing

Simon Thornber, GSK
Cary O'Donnell, AZ
Quan Yang, Novartis
Monica Arenz, Novartis

Steering Group:-
Ashley George, GSK
Tom Flores, GSK
Martyn Wilkins
Patrick Warren

VSI

Lee Harland, Christopher Larminie,
Ian Dix, Wendy Filsell, OBO PRO