“ELIXIR and Open Data”

View from an ELIXIR Node”

Barend Mons
Prof. Biosemantics, LUMC, Scientific director NBIC,
Head of ELIXIR Node

18 December 2013
ELIXIR Launch, Brussels

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org
Outline

• The Dutch Node
• Data in the eScience era: Pattern Recognition and Excavation
• Data Collection, Archiving and: Reduction
• Why is ELIXIR important for Open Data?
• What are the needs of clinical institutes and Industry: Open and Managed Data
• Why training as part of ELIXIR?
Data interoperability and exchange

Compute and storage infrastructure services

Training & Education

Nodes and a Hub in NL (DTL) > Node in >>>>>>>>
The Data cycle in eScience

Data Stewardship covers the entire data cycle.
The Data Challenge

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady
- DNA sequence data is **doubling every 6-8 months** over the last 3 years and looks to continue for this decade

Proper Data stewardship and analysis may be THE limiting factor in eScience
Simplified eScience

The Goal is Knowledge Discovery, not Data Collection
Pattern Recognition in Open Data and detailed Excavation should be separated
How do we discover patterns in ‘Ridiculograms’?

The Explicitome:

$10^{14}$

Individual explicit associations
The Semantic Web approach to interoperability

The Unique Explicititome: $10^{11}$ Cardinal Assertions
We publish about less than a million LS concepts

10^6 concept clusters (Knowlets)
Zipping the Explicitome

$10^{14}$

Individual explicit associations

$10^{11}$

CA’s

$5 \times 10^5$

Knowlets

$\approx 99.999996\%$ reduction of infoburden
In silico knowledge discovery for the millions..

Reasoning takes place on aggregated and zipped data
The Implicitome

> 1 M hypotheses hidden in the implicitome
The Goal is Knowledge Discovery, not Data Collection
The Role of ELIXIR: Open versus Managed

- Clinical Data sets
- Public Data sets
- Private Data sets
- ELIXIR Data sets

Tools and standards for interoperability

Data Collection(s) Anywhere, everywhere
For Big Data to become huge, however, there are still hurdles to leap. For one thing, the tools to analyse data are not yet good enough. And people with the skills to analyse data are scarce and will become scarcer. By 2018 there will be a “talent gap” of between 140,000 and 190,000 people.
Vision (personal, not necessarily ELIXIR)

- Data Collections are not a goal in themselves
- They ultimately serve **Knowledge Discovery**
- E-datastewardship in a time of plenty is thus also **data zipping**
- E-datastewardship should address the **entire data cycle**
- ELIXIR is more about a trusted partner for the ‘Tools&Rules’ for data stewardship and interoperability than about data archiving.
- Interoperability is for **both** humans and computers.
- **Open data needs to be talking to ‘closed data’**
- **Data experts** need the place in the egosystem they deserve
SHARED knowledge is Double Knowledge

‘Knowledge is like love: it multiplies when shared’