Semantic Web Technology for Health Care and Life Science

David Hansen
Project Leader, EHRC

EHEALTH RESEARCH CENTRE

CSIRO

Outline

- E-Health, Health Data and Health Informatics
- Health care and life science informatics
- Semantic Web for Health Care and Life Science
- The Bench to Bedside challenge

e-Health Research Centre

www.e-hrc.net

- $15 million JV between CSIRO and Queensland Government
- Undertake applied research and conduct clinical trials in conjunction with State Health agencies
- Research Program : “Improving Disease Management – providing information and knowledge where it is needed, when it is needed and in the form it is needed to support clients, patients and clinicians.”

Queensland Government

CSIRO

eHEALTH RESEARCH CENTRE

CSIRO

E-Health and Health Data and Health Informatics

A Perspective on the Scope of e-Health

- Contemporary health care has adopted evidence-based medicine delivered by ***multi-disciplinary, multi-party*** health care teams in a ***patient-centred*** approach
  - Synonymous with the information age
  - e-Health encompasses the broad application of Information and Communication Technologies in support of health care needs
  - The main e-Health domains of activity are:
    - Health Information Systems (data and software tools)
    - Health Services Delivery (work practices and processes)

Current Major Trends in e-Health

- Universal electronic health records and data integration
- Health information portals and patient empowerment
- Clinical decision support systems and individual care
- Remote patient monitoring (telecare) and telemedicine
- Medical imaging and sensing with new modalities
- Computer-assisted procedures and virtual surgery
- Genome and protein analysis for phenotype diagnostics
- Drug and therapy development by translational medicine
- etc. . .

Global programs in e-Health

- UK - Connecting for health
  - Patient access to their electronic health records
  - Connecting 30,000 GPs to 300 hospitals throughout the UK
- USA
  - Federal initiatives to encourage electronic health records for patients
  - Institute/insurance company based
- Australia – NEHTA
  - Provider and Consumer Ids
  - Terminology's
    - SNOMED CT

→ Primarily about patient records rather than data for research purposes

Need and Benefits of Integrating the Data

- Currently patient data resides across numerous different databases which are unconnected
  - Different information systems and reporting systems
  - Government vs Hospital vs GP vs Allied health systems

- Health care improvement opportunities flow from using this data together
  - higher levels of patient care due to fuller information
  - extension of evidence-based practice
  - better planning or decision making for specific cases

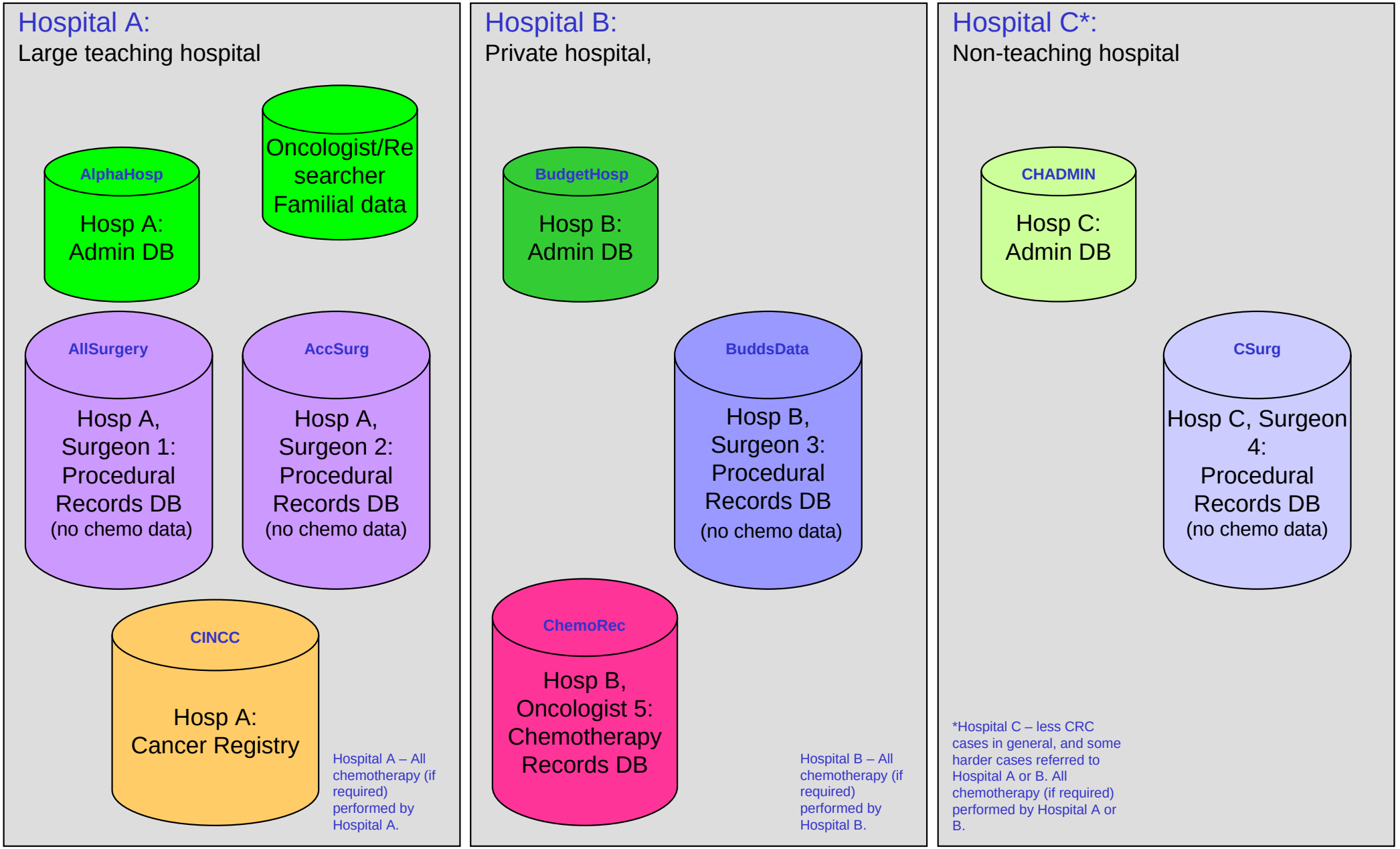
Health care and life science informatics

Health care data sources

- National level
  - Medicare – cost codes identify treatments
  - PBS - Pharmaceuticals Benefits Scheme
- State data
  - Health department hospital admissions data
  - State based disease-specific data collections
  - Pathology reports
  - Radiotherapy reports
  - Registries
- Hospital data
  - Hospital Information system
  - Hospital pharmaceuticals database
- Hospital unit data
  - Clinical information systems
  - Unit specific data sources
- Clinical area data sources
  - Clinician based data sources
- Genomic data
  - molecular sequences
  - protein structures
  - SNPs and haplotypes,
  - expression microarrays
  - Bio/medical literature
- Environmental data
  - Particulate data
- Remote sensor data
  - ECG, movement monitoring

Example Scenario for Colorectal Cancer Research

www.e-hrc.net



Knowledge from integrating and analysing health data

www.e-hrc.net

- Quality and safety of patient care
  - adherence to clinical guidelines
  - comparison across hospitals
- Treatment outcomes
  - To support evidence based care
- Screening sensitivity
  - Determining better information for patients regarding likely screening outcomes by age and sex and familial risk
  - Analysing optimum screening schedules by familial risk
- Diagnostic sensitivity
  - Sensitivity and specificity of the Faecal Occult Blood Test (FOBT)
- Surgical Outcomes
  - By factors such as smoking history, adjuvant therapy or diabetes status

On a larger scale:

- Integrating cancer screening information across state boundaries to obtain a larger cohort for research and analysis
  - Familial risk and screening outcomes can be captured in vastly different formats across states, resulting in difficulties in mapping to a common standard

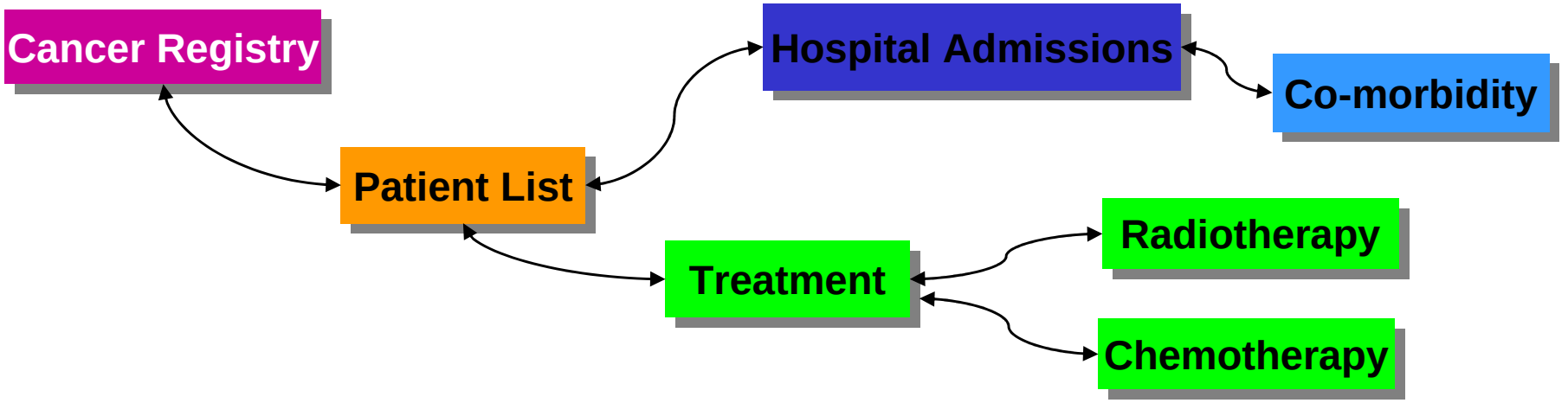
Typical health informatics infrastructure

- Patient Data spread across several databases
  - Hospital administration and clinical databases
  - Pathology and Pharmaceuticals databases
  - Various information systems and legacy infrastructure
  - Collected for a variety of reasons – including administrative
- No common person identifier
  - New identifying number or ID per institution or service provider
- Privacy and Security
  - Patient concerns, Legislative requirements, Data ownership concerns
- No way to easily manage access to data in multiple databases
  - Involves significant time and manual handling of data between computer systems
- Data quality and consistency
  - Data entry errors
  - Non-standard coding and formats
- Ad-hoc Analysis and Reporting
  - To meet requirements from state and federal health departments

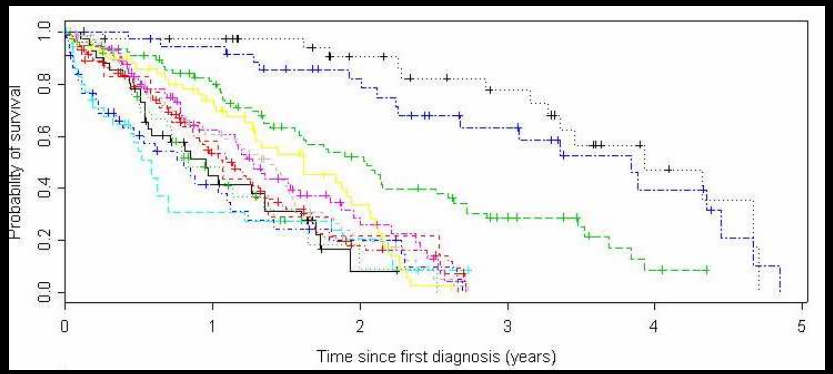
Too much data, too many formats, too little time

Integration Supports Enquiry

www.e-hrc.net

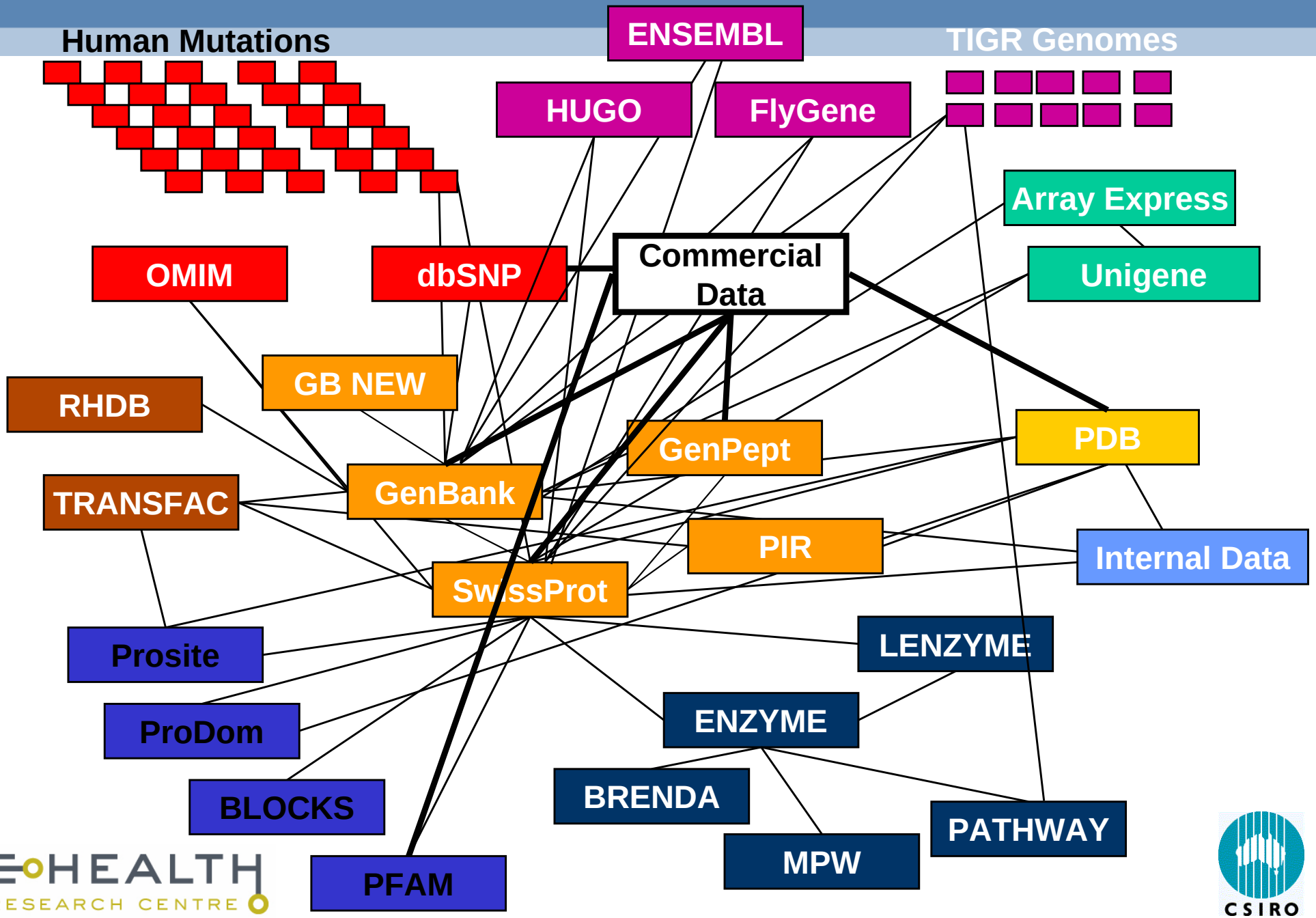


Show me survival data for *CRC patients* diagnosed in 2001 treated with radiotherapy or chemotherapy by co-morbidity



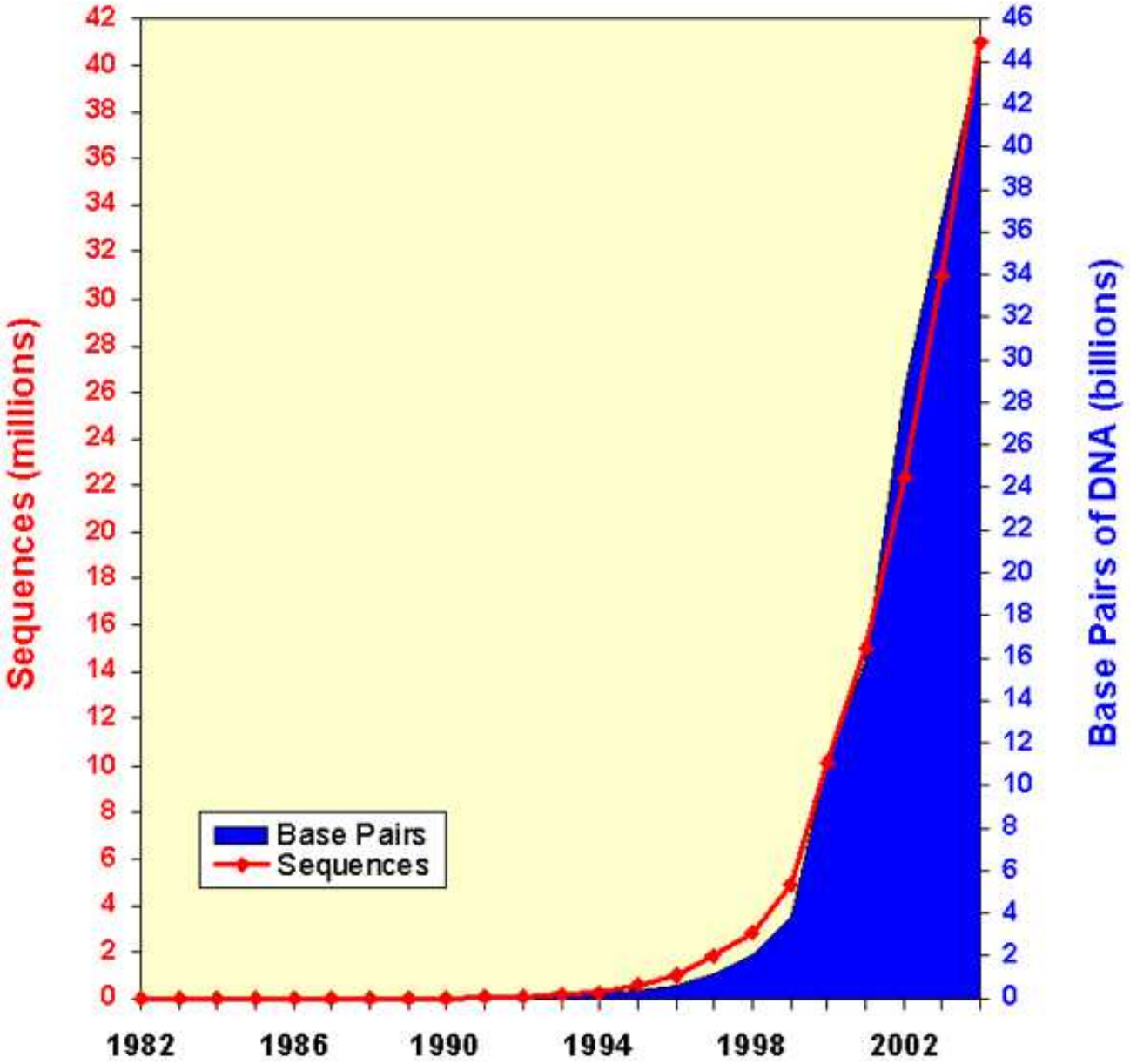
What about life science?

www.e-hrc.net



Growth of GenBank

(1982 - 2004)

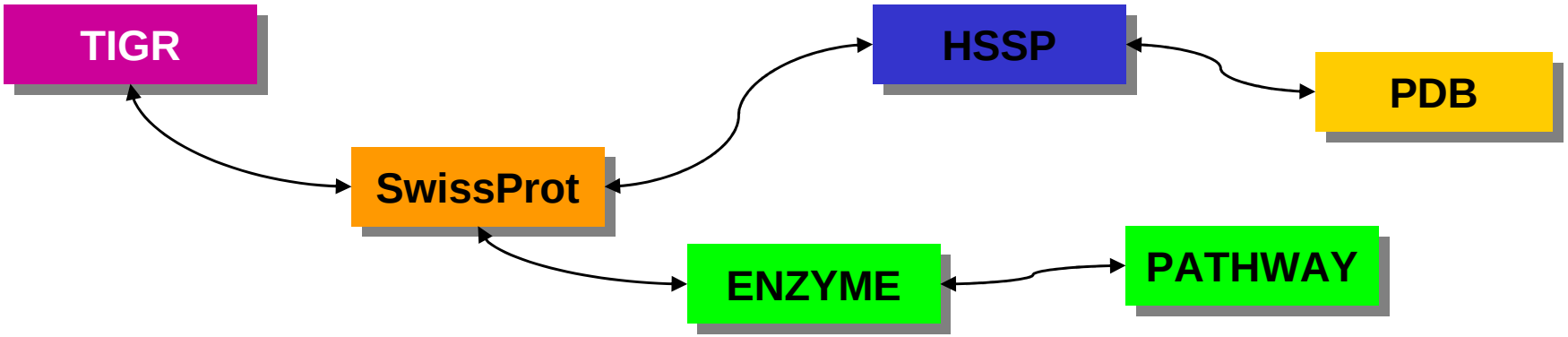


Typical bioinformatics infrastructure

- Flat files and relational databases store ever increasing data size and formats
- Perl scripts used extensively to parse GenBank\EMBL\SWISS_PROT etc
- BLAST and FASTA used for analysis
  - EMBOSS is used increasingly
- Some internal tools for analysis and visualization
- Third party data, e.g.. Incyte
- Internally generated data in ORACLE (and other RDBs) or XML format

Too much data, too many formats, too little time

Integration Supports Enquiry



All *H. pylori* genes
encoding membrane bound proteins
involved in glucose metabolism
and with a homologue of
known 3D structure with
resolution better 2Å

Semantic Web for Health Care and Life Science

W3C HCLSIG Members

- Health Care, Clinical, and Life Science Consortia
- Research Institutes and Centers
- Pharmaceuticals and Biotechnology Companies
- IT Solution Vendors
- Government Agencies

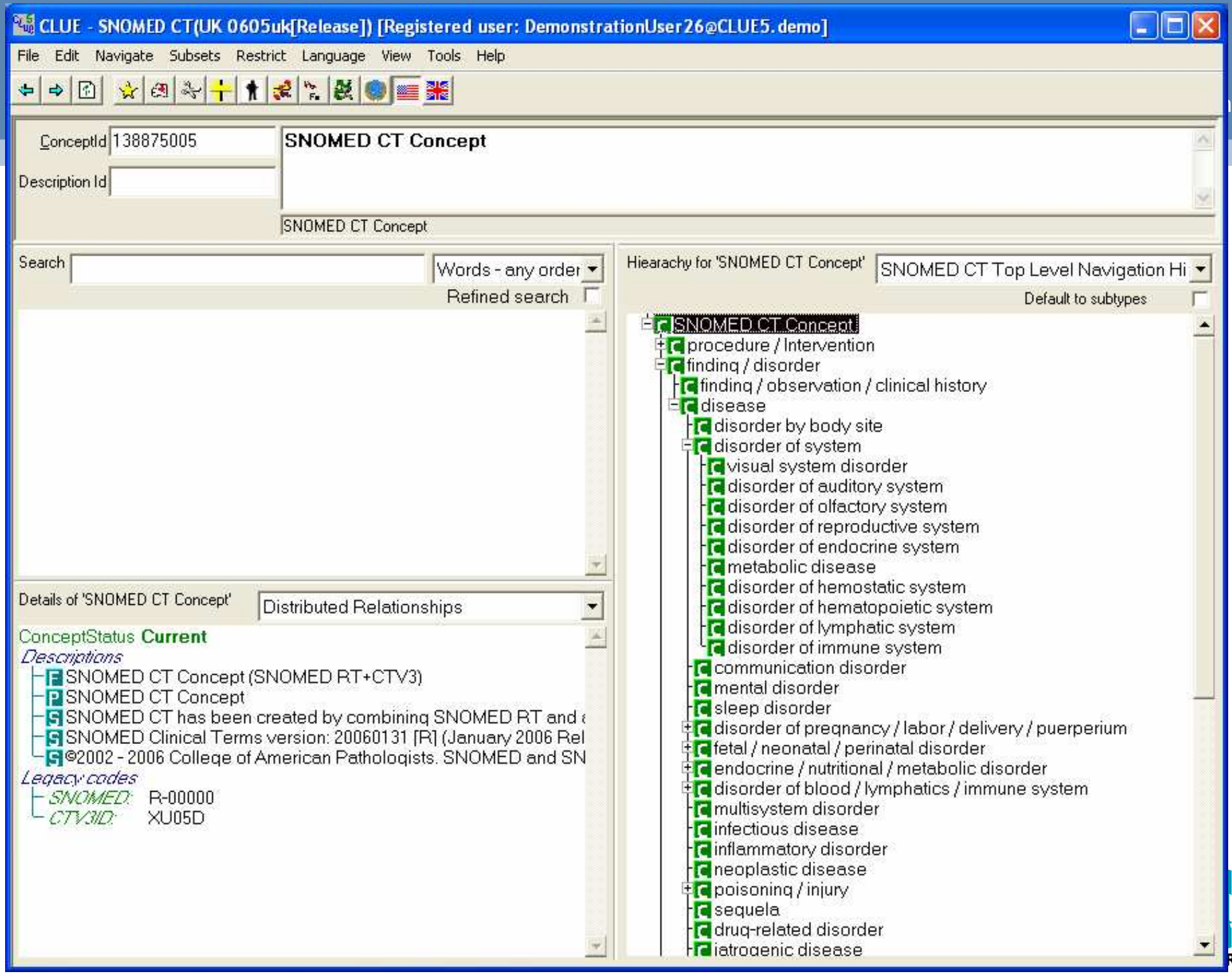
W3C Semantic Web Health Care and Life Science Interest Group

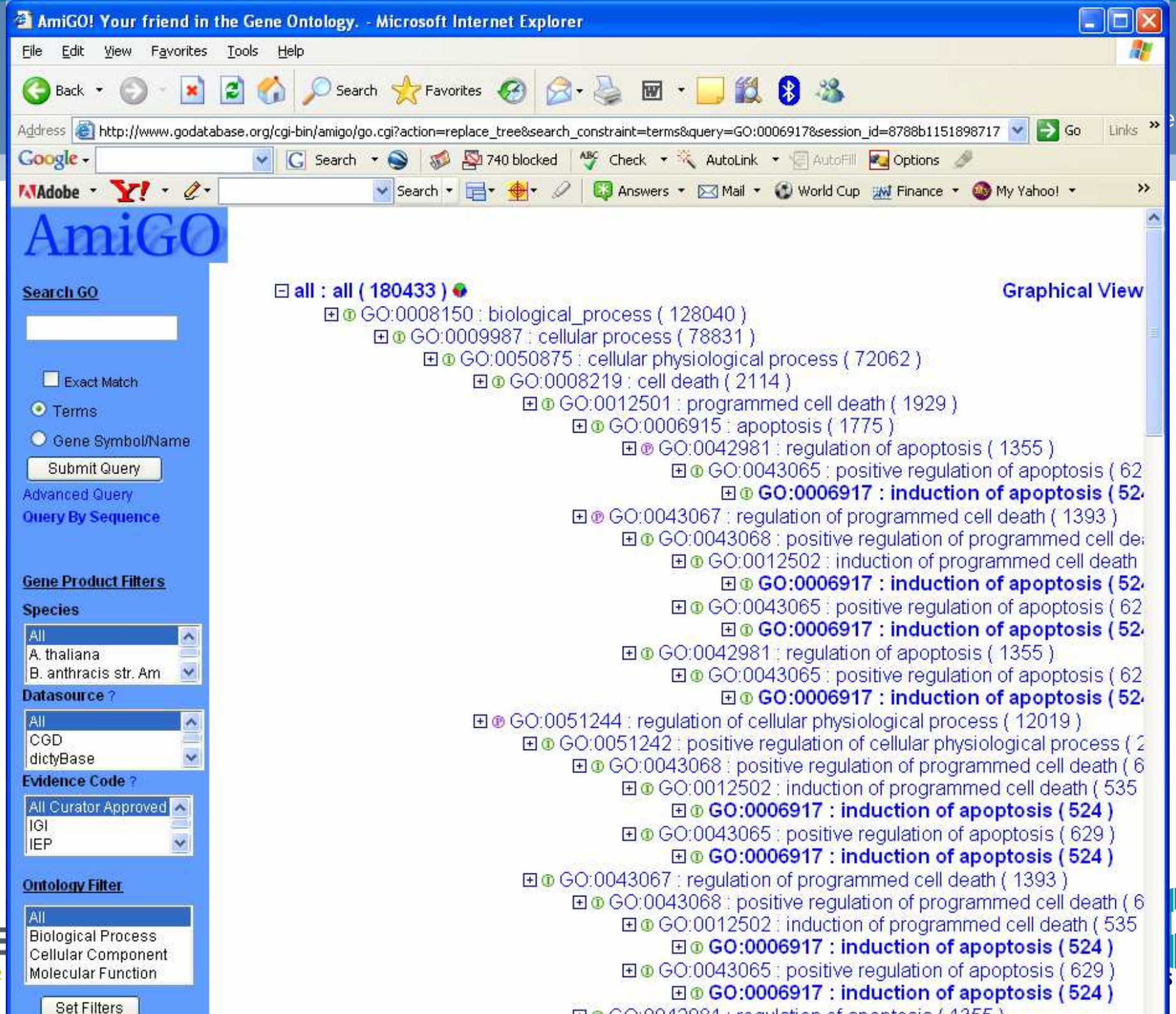
www.e-hrc.net

- Develop use cases for Semantic Web technology, rather than develop standards
  - Core vocabularies and ontologies to support cross-community data integration and collaborative efforts
  - Guidelines and Best Practices for Resource Identification to support integrity and version control
  - Better integration of Scientific Publication with people, data, software, publications, and clinical trials

Core Vocabularies

- Health care/Clinical Terminology
  - SNOMED CT
    - 400 thousand concepts, almost 1 million relationships
  - UMLS
- Life Science
  - The Gene Ontology (GO)
  - MGED
  - SBML
  - Over 40 other publicly available ontologies
- Upper Ontology work
  - OBO – open biomedical ontologies project @ EBI


CLUE - SNOMED CT(UK 0605uk[Release]) [Registered user: DemonstrationUser26@CLUE5.demo]
File Edit Navigate Subsets Restrict Language View Tools Help
ConceptId 138875005
SNOMED CT Concept
Description Id
SNOMED CT Concept
Search
Words - any order
Refined search
Hierarchy for 'SNOMED CT Concept'
SNOMED CT Top Level Navigation Hi
Default to subtypes
SNOMED CT Concept
procedure / intervention
finding / disorder
finding / observation / clinical history
disease
disorder by body site
disorder of system
visual system disorder
disorder of auditory system
disorder of olfactory system
disorder of reproductive system
disorder of endocrine system
metabolic disease
disorder of hemostatic system
disorder of hematopoietic system
disorder of lymphatic system
disorder of immune system
communication disorder
mental disorder
sleep disorder
disorder of pregnancy / labor / delivery / puerperium
fetal / neonatal / perinatal disorder
endocrine / nutritional / metabolic disorder
disorder of blood / lymphatics / immune system
multisystem disorder
infectious disease
inflammatory disorder
neoplastic disease
poisoning / injury
sequela
drug-related disorder
iatrogenic disease
Details of 'SNOMED CT Concept'
Distributed Relationships
ConceptStatus Current
Descriptions
SNOMED CT Concept (SNOMED RT+CTV3)
SNOMED CT Concept
SNOMED CT has been created by combining SNOMED RT and
SNOMED Clinical Terms version: 20060131 [R] (January 2006 Rel
©2002 - 2006 College of American Pathologists. SNOMED and SN
Legacy codes
SNOMED: R-00000
CTV3ID: XU05D


AmiGO! Your friend in the Gene Ontology. - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.godatabase.org/cgi-bin/amigo/go.cgi?action=replace_tree&search_constraint=terms&query=GO:0006917&session_id=8788b1151898717
Go Links
Google Search 740 blocked Check AutoLink AutoFill Options
Adobe Search Answers Mail World Cup Finance My Yahoo!
AmiGO
Search GO
Exact Match
Terms
Gene Symbol/Name
Submit Query
Advanced Query
Query By Sequence
Gene Product Filters
Species
All
A. thaliana
B. anthracis str. Am
Datasource ?
All
CGD
dictyBase
Evidence Code ?
All Curator Approved
IGI
IEP
Ontology Filter
All
Biological Process
Cellular Component
Molecular Function
Set Filters
all : all (180433)
Graphical View
GO:0008150 : biological_process (128040)
GO:0009987 : cellular process (78831)
GO:0050875 : cellular physiological process (72062)
GO:0008219 : cell death (2114)
GO:0012501 : programmed cell death (1929)
GO:0006915 : apoptosis (1775)
GO:0042981 : regulation of apoptosis (1355)
GO:0043065 : positive regulation of apoptosis (62
GO:0006917 : induction of apoptosis (52
GO:0043067 : regulation of programmed cell death (1393)
GO:0043068 : positive regulation of programmed cell de
GO:0012502 : induction of programmed cell death
GO:0006917 : induction of apoptosis (52
GO:0043065 : positive regulation of apoptosis (62
GO:0006917 : induction of apoptosis (52
GO:0042981 : regulation of apoptosis (1355)
GO:0043065 : positive regulation of apoptosis (62
GO:0006917 : induction of apoptosis (52
GO:0051244 : regulation of cellular physiological process (12019)
GO:0051242 : positive regulation of cellular physiological process (2
GO:0043068 : positive regulation of programmed cell death (6
GO:0012502 : induction of programmed cell death (535
GO:0006917 : induction of apoptosis (524)
GO:0043065 : positive regulation of apoptosis (629)
GO:0006917 : induction of apoptosis (524)
GO:0043067 : regulation of programmed cell death (1393)
GO:0043068 : positive regulation of programmed cell death (6
GO:0012502 : induction of programmed cell death (535
GO:0006917 : induction of apoptosis (524)
GO:0043065 : positive regulation of apoptosis (629)
GO:0006917 : induction of apoptosis (524)

Task Groups

- BIORDF (Structured Data to RDF)
- Knowledge Life Cycle
- Ontologies Working Group
- Adaptive Healthcare Protocols and Pathways
- ROI Analysis within HCLS

Bench to Bedside Challenge

Issues for Health Care and Life Science

- Growth of knowledge base
  - Medical literature doubling every 19 years, but every 22 months for AIDS care
- Clinical decision support systems
  - Will need to support
    - Increasingly complex relationships between data
    - Exponential growth of known factors
    - New diagnostics, particularly molecular becoming available daily
- Drug discovery
  - Reduce the risk, duration and cost of
    - drug discovery by using all available data
    - clinical trials

→ Speed the translational medicine life cycle