Grid enabling 'real' science and engineering?

David Abramson

Monash e-Science and Grid Engineering Lab (MESSAGE Lab)
Faculty of Information Technology
Acting Director: Monash e-Research Centre
ARC Professorial Fellow

Contents

• The changing face of science
• Grid testbeds: The Pragma testbed
• Software tools for "real" science
  – The MESSAGE Lab approach
• Some successful projects
  – Environmental science,
  – Systems biology,
  – Chemistry
  – Physics
  – Engineering
• Conclusions
THE CHANGING FACE OF SCIENCE

e-Science

- Pre-Internet
  - Theorize &/or experiment, alone or in small teams; publish paper
- Post-Internet
  - Construct and mine large databases of observational or simulation data
  - Develop simulations & analyses
  - Access specialized devices remotely
  - Exchange information within distributed multidisciplinary teams

"Grids are not just communities of computers, but communities of researchers, of people."
— Peter Arzberger, UCSD
Typical Grid Applications

- Characteristics
  - High Performance Computation
  - Distributed infrastructure
  - Instruments are first class resources
  - Lots of data
  - Not just bigger - fundamentally different

- Some examples
  - In-silico biology (See MyGrid)
  - Earthquake simulation
  - Virtual observatory
  - High energy physics
  - Medical applications
  - Environmental applications
The Grid

- Infrastructure ("middleware" & "services") for establishing, managing, and evolving multi-organizational federations
  - Dynamic, autonomous, domain independent
  - On-demand, ubiquitous access to computing, data, and services
- Mechanisms for creating and managing workflow within such federations
  - New capabilities constructed dynamically and transparently from distributed services
  - Service-oriented, virtualization

Source: Ian Foster
What is a Grid?

• Three key criteria
  – Coordinates distributed resources ...
  – using standard, open, general-purpose protocols and interfaces ...
  – to deliver non-trivial qualities of service.

• What is not a Grid?
  – A cluster, a network attached storage device, a scientific instrument, a network, etc.
  – Each may be an important component of a Grid, but by itself does not constitute a Grid.

Source: Ian Foster

The (Power) Grid: On-Demand Access to Electricity

Source: Ian Foster
By analogy, some challenges

Voltage 110 – 220 – 240
Frequency 50 – 60 Hz.

THE PRAGMA TESTBED: A GRASS ROOTS GRID
What have we learned?

- Faults (hardware, software and network) are normal
- Testbed supports a rich variety of application needs
- Testbed is large,
  - resource availability, network bandwidth and contention, distributed resource monitoring, and global name spaces.
- Coordination of resources, monitoring, inter-operable software deployments, and resource/user policies complicated because applications may have different requirements for middleware and resource allocation.
- Users and administrators located across 12 time zones.
  - challenge in putting together a team with very diverse cultural backgrounds.
- Countries with different levels of network speed, quality, and system resources.
- Strong feedback between application and middleware (e.g. Ninf-G, Nimrod and Gfarm).

Supporting the software life-cycle

Applications Development Test & Debug Deploy & Build Execution

Environmental Sciences Life & Pharmaceutical Sciences Geo Sciences

Development Test/Debug Deploy Execution

Upper Middleware /Tools
Nimrod KE/IO/IE Nimrod/K Kepler Eclipse Eclipse VB Guard DisANT Nimrod Portal & WB GridLab NePFiles Excel Active Sheets

Lower Middleware
Oracle DB Debug Deploy Motor REMUS Active Data SSH

Platform Infrastructure
Unix Windows JVM TCP/IP MPI .Net Runtime Web Services VPN SSH

Nimrod Development Cycle
Prepare Jobs using Portal
Jobs Scheduled Executed Dynamically
Sent to available machines
Results displayed & interpreted
Ionisation Chamber Design
Lew Kotler, ARPANSA

Direction of photons

Field shaping electrodes

Front wall

Guard electrode

Polystyrene Insulator

Aluminium

Teflon Insulator

Collecting electrode

Ion collection volume
(10 mm diameter x 21 mm length)

EGS4 Simulation of response of Aluminium Cavity
Chamber to 60Co photons
Radiotherapy planning
Giddy, Chin, Lewis, Welsh e-Science Centre, UK

BEAMnrc

EGS

DOSXYZnrc

www.utsouthwestern.edu/.../270177Synergy5.2.bmp

Outcomes

CONVOLUTION
/SUPERPOSITION

MONTE CARLO

SmartPET - A Compton Camera
Toby Beveridge, Monash University

A SmartPET Detector
• Large Volume - 20 x 60 x 60 mm³
• Operating Range 0.1 – 2 MeV
• Detector resolution depends on Pulse Shape Analysis
• Multi-parameter space is difficult to characterise, and optimise, analytically
• Monte-Carlo solutions such as GEANT4 are computationally expensive

A Compton Camera
• Extensive FoV
• Multi-resolution Data
• Angular precision depends on detector resolution

Outcomes

For a Single Trial
242 point-source locations (11² field over 2 orthogonal planes)
5 energies (between 140 keV and 1000 keV)
2 different detection conditions
242 x 5 x 2 x (20 mins per run) = 806 hours

Each pixel (at a particular incident energy) was assigned a separate job
At each point the resolution matrix could be calculated
ENVIRONMENT

Climate Studies
Lynch, Abramson, Görgen, Beringer, Uotila, Monash University

• Extensive savanna eco-systems in northern Australia
• Changing fire regime
• Fires lead to abrupt changes in surface properties
  – Surface energy budgets
  – Partitioning of convective fluxes
  – Increased soil heat flux
  – Modified surface-atmosphere coupling
• Sensitivity study: do the fire's effects on atmospheric processes lead to changes in highly variable precipitation regime of Australian Monsoon?
• Many potential impacts (e.g. agricultural productivity)
Outcomes

A Workshop On Earth System Models of Intermediate Complexity
28-29 March 2006 at the Bureau of Meteorology Research Centre, Melbourne

SYSTEMS BIOLOGY
Heart disease still leading cause of death

- Understanding the underlying physiological mechanisms is cheaper and faster when experimental studies are performed together with mathematical models & computer simulations

- Studying pathologies

- Developing & Testing drugs

Cardiac Modelling

Sher, Gavaghan, Hinch, Noble, Oxford University

- Based on experimental data, mathematical models have been developed
  - ODE’s
  - Initial conditions
  - Ion movement in single cells

Cardiac Modeling

- Shannon et al. model, 2004
Studying ionic models
Anna Sher, Oxford

• Examine the effect of various parameters on Ca2+-induced Ca2+ release and on shape of the action potential
• Fit simulated to experimental data
• Identify parameter(s) that are critical to distinguish Ca2+ dynamics within various species

Outcomes

• Single cell ionic models allow us to study:
  – Whole cell currents during an action potential (AP)
  – Currents in response to voltage-clamp stimuli
  – Dynamics of ions such as Ca2+ and Na+
  – Force-frequency relationship
  – etc.
Cardiac Modelling
Dederko, Nevo, Altshuler, Wu, McCulloch, Mihaylova, Kerckhoffs, UCSD

Active Potential for Epicardium Cells under Stretch Activated Currents

CHEMISTRY
Quantum Chemistry
Wibke Sudholt, Univ Zurich

\[ U_{\text{eff}}(r) = A_1 \exp(-B_1 r^3) + A_2 \exp(-B_2 r^3) \]

Drug docking pipeline
Baldridge, Amoreira, Univ Zurich, Berstis, Kondrick, UCSD

- Goal is to minimize the free binding energy
- Use Quantum calculations for more realism
Flame Kernel Growth in Turbulent Flows
Tom Dunstan, Karl Jenkins, Cranfield University

Laminar Flame solutions
Laminar simulations are carried out for verification of the code and adjustment of the initial reaction parameters.

Fig 1. Non-dimensional flame speed. time step = 3x10^{-5} s

Fig 2. Progress variable C (top) and Reaction Rate (bottom) at progressive time steps through centre of domain.

NIMROD optimisation results
Mass Fraction Burned describes global state of reaction (from 0 in unburned state to 1 in fully burned).

Fig 3. Mass fraction burned for varying values of pre-exponential factor B* and heat release parameter.
Turbulent Flame propagation

Conclusions

• Significant benefits for science and engineering
• Users want to focus on their research
• Tools need to be powerful and hide the complexity of the underlying platforms
• Real studies improve tools
• Application studies are significant component of the research

www.csse.monash.edu.au/~davida
Grid enabling 'real' science and engineering?

David Abramson
Monash e-Science and Grid Engineering Lab (MESSAGE Lab)
Faculty of Information Technology
Acting Director: Monash e-Research Centre
ARC Professorial Fellow

Contents

- The changing face of science
- Grid testbeds: The Pragma testbed
- Software tools for "real" science
  - The MESSAGE Lab approach
- Some successful projects
  - Environmental science,
  - Systems biology,
  - Chemistry
  - Physics
  - Engineering
- Conclusions
THE CHANGING FACE OF SCIENCE

e-Science

• Pre-Internet
  – Theorize &/or experiment, alone
     or in small teams; publish paper

• Post-Internet
  – Construct and mine large databases of observational or
    simulation data
  – Develop simulations & analyses
  – Access specialized devices remotely
  – Exchange information within distributed multidisciplinary teams

“Grids are not just communities of computers, but communities of researchers, of people.” — Peter Arzberger, UCSD
Typical Grid Applications

- Characteristics
  - High Performance Computation
  - Distributed infrastructure
  - Instruments are first class resources
  - Lots of data
  - Not just bigger - fundamentally different

- Some examples
  - In-silico biology (See MyGrid)
  - Earthquake simulation
  - Virtual observatory
  - High energy physics
  - Medical applications
  - Environmental applications
The Grid

- Infrastructure ("middleware" & "services") for establishing, managing, and evolving multi-organizational federations
  - Dynamic, autonomous, domain independent
  - On-demand, ubiquitous access to computing, data, and services
- Mechanisms for creating and managing workflow within such federations
  - New capabilities constructed dynamically and transparently from distributed services
  - Service-oriented, virtualization

Source: Ian Foster
What is a Grid?

• Three key criteria
  – Coordinates distributed resources ...
  – using standard, open, general-purpose protocols and interfaces ...
  – to deliver non-trivial qualities of service.
• What is not a Grid?
  – A cluster, a network attached storage device, a scientific instrument, a network, etc.
  – Each may be an important component of a Grid, but by itself does not constitute a Grid.

Source: Ian Foster

The (Power) Grid: On-Demand Access to Electricity

Source: Ian Foster
By analogy, some challenges

Voltage 110 – 220 – 240
Frequency 50 – 60 Hz.

THE PRAGMA TESTBED: A GRASS ROOTS GRID
32 Clusters from 29 institutions in 14 countries/regions (+7 in preparation) 7 gfarm sites

Source Cindy Zheng
What have we learned?

- Faults (hardware, software and network) are normal
- Testbed supports a rich variety of application needs
- Testbed is large, resource availability, network bandwidth and contention, distributed resource monitoring, and global name spaces.
- Coordination of resources, monitoring, inter-operable software deployments, and resource/user policies complicated because applications may have different requirements for middleware and resource allocation.
- Users and administrators located across 12 time zones, challenge in putting together a team with very diverse cultural backgrounds.
- Countries with different levels of network speed, quality, and system resources.
- Strong feedback between application and middleware (e.g. Ninf-G, Nimrod and Gfarm).

Supporting the software life-cycle

Applications
- Environmental Sciences
- Life & Pharmaceutical Sciences
- Geo Sciences

Development
- Nimrod/G
- Eclipse
- VB
- DaliANT
- GridLab

Test/Debug
- DistANT
- Kepler
- Nimrod/K
- Eclipse
- Guard

Deploy
- Nimrod Portal & WB
- NetFiles
- Excel

Execution
- Active Sheets
- Excel

Platform Infrastructure
- Unix
- Windows
- JVM
- TCP/IP
- MPI
- Active Data

Upper Middleware /Tools
- Nimrod/G
- KEPLER
- DistANT
- GridLab

Lower Middleware
- Deploy
- GridLab
- NetFiles

Middleware
- Java
- XML
- TCP/IP
- MPI
- .Net Runtime

Supporting the software life-cycle

Nimrod Development Cycle

Prepare Jobs using Portal

Sent to available machines

Jobs Scheduled Executed Dynamically

Results displayed & interpreted
Ionisation Chamber Design
Lew Kotler, ARPANSA

- Front wall
- Guard electrode
- Collecting electrode
- Ion collection volume (10 mm diameter x 21 mm length)
- Direction of photons
- Polystyrene Insulator
- Aluminium
- Teflon Insulator

EGS4 Simulation of response of Aluminium Cavity Chamber to $^{60}$Co photons

Inset: EGS4 calculation fitted to sum of exponentials.

Graph: Thickness of front cap vs. Ion pairs per incident photon.
Radiotherapy planning
Giddy, Chin, Lewis, Welsh e-Science Centre, UK

BEAMnrc
EGS
DOSXYZnrc

Outcomes

CONVOLUTION
/SUPERPOSITION

MONTE CARLO

SmartPET - A Compton Camera
Toby Beveridge, Monash University

A SmartPET Detector
• Large Volume - 20 x 60 x 60 mm³
• Operating Range 0.1 – 2 MeV
• Detector resolution depends on Pulse Shape Analysis

A Compton Camera
• Extensive FoV
• Multi-resolution Data
• Angular precision depends on detector resolution

• Multi-parameter space is difficult to characterise, and optimise, analytically
• Monte-Carlo solutions such as GEANT4 are computationally expensive

Outcomes

Each pixel (at a particular incident energy) was assigned a separate job
At each point the resolution matrix could be calculated

For a Single Trial
242 point-source locations (11² field over 2 orthogonal planes)
5 energies (between 140 keV and 1000 keV)
2 different detection conditions
242 x 5 x 2 x (20 mins per run) = 806 hours
Climate Studies
Lynch, Abramson, Görgen, Beringer, Uotila, Monash University

- Extensive savanna eco-systems in northern Australia
- Changing fire regime
- Fires lead to abrupt changes in surface properties:
  - Surface energy budgets
  - Partitioning of convective fluxes
  - Increased soil heat flux
  - Modified surface-atmosphere coupling
- Sensitivity study: do the fire's effects on atmospheric processes lead to changes in highly variable precipitation regime of Australian Monsoon?
- Many potential impacts (e.g., agricultural productivity)
Outcomes

A Workshop On Earth System Models of Intermediate Complexity
28-29 March 2006 at the Bureau of Meteorology Research Centre, Melbourne

SYSTEMS BIOLOGY
Heart disease still leading cause of death

Understanding the underlying physiological mechanisms is cheaper and faster when experimental studies are performed together with mathematical models & computer simulations

Studying pathologies

Developing & Testing drugs

Cardiac Modelling

Sher, Gavaghan, Hinch, Noble, Oxford University

Cardiac Modeling

Based on experimental data, mathematical models have been developed
– ODE’s
– Initial conditions
– Ion movement in single cells
Studying ionic models
Anna Sher, Oxford

• Examine the effect of various parameters on Ca2+-induced Ca2+ release and on shape of the action potential
• Fit simulated to experimental data
• Identify parameter(s) that are critical to distinguish Ca2+ dynamics within various species

Outcomes

• Single cell ionic models allow us to study:
  – Whole cell currents during an action potential (AP)
  – Currents in response to voltage-clamp stimuli
  – Dynamics of ions such as Ca2+ and Na+
  – Force-frequency relationship
  – etc.
Cardiac Modelling
Dederko, Nevo, Altshuler, Wu, Mcculloch, Mihaylova, Kerckhoffs, UCSD

Action Potential for Endocardium Cells under Stretch Activated Currents

VA/Time (s)
Drug docking pipeline

Baldridge, Amoreira, Univ Zurich, Berstis, Kondrick, UCSD

- Goal is to minimize the free binding energy
- Use Quantum calculations for more realism
Flame Kernel Growth in Turbulent Flows
Tom Dunstan, Karl Jenkins, Cranfield University

Laminar Flame solutions
Laminar simulations are carried out for verification of the code and adjustment of the initial reaction parameters.

Fig 1. Non-dimensional flame speed, time step = $3 \times 10^{-5}$ s

Fig 2. Progress variable C (top) and Reaction Rate (bottom) at progressive time steps through centre of domain.

NIMROD optimisation results
Mass Fraction Burned describes global state of reaction (from 0 in unburned state to 1 in fully burned).

Fig 3. Mass fraction burned for varying values of pre-exponential factor $A^*$ and heat release parameter.
Conclusions

- Significant benefits for science and engineering
- Users want to focus on their research
- Tools need to be powerful and hide the complexity of the underlying platforms
- Real studies improve tools
- Application studies are significant component of the research

www.csse.monash.edu.au/~davida
Data Linking and Integration for Health Applications

eResearch Australasia, Brisbane, 27 June 2007

Dr Anthony Maeder
Research Director
E-Health Research Centre / CSIRO ICT Centre
Brisbane, Queensland, Australia

Overview

- E-Health and Health Data
- The HDI Software Tool
- Current Projects
- Future Directions

http://ict.csiro.au/
E-Health and Health Data

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
- 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)
Australian Health Systems Scene

- State-based systems, with different structures for management of hospitals and community health
- National and local organisation of private providers
- Australian Government provides financial resource
- Many independent legacy software systems: specialised, non-interoperable, unsupported
- Safety, Quality and Efficiency issues driving reform of work practices and wider sharing of information
- Key problem is lack of universal health identifier: National e-Health Transition Authority aims at this

Need and Benefits of Data Linking

- Currently patient data resides across numerous different databases which are unconnected and owned separately
  - Different information systems and reporting systems
  - Government vs Hospital vs GP vs Allied health systems
- Health care improvement opportunities flow from linking this data
  - higher levels of patient care due to fuller information
  - extension of evidence-based practice
  - better planning or decision making for specific cases
  - improvements to training and education, safety and quality
Privacy Issues

- **National Privacy Principles**
  - Use of health data for treatment purposes
  - Secondary use of data for research purposes

- **State Acts related to Health Data:**
  - Health Records and Information Privacy Act (2002) - NSW
  - Health Records Act (2001) – Victoria
  - Many others

- **Organisation principles**
  - Other policies applicable site by site
  - Access to data governed by ethics compliance
Health Data Sources

- **National level**
  - Medicare – cost codes identify treatments
  - PBS - Pharmaceuticals Benefits Scheme

- **State level**
  - Health department hospital admissions data
  - State based disease-specific data collections
  - Pathology reports & results
  - Radiotherapy reports
  - Radiology reports & images
  - Registries

- **Hospital level**
  - Hospital Information system
  - Hospital pharmaceuticals database

- **Hospital units**
  - Clinical information systems
  - Unit specific data sources

- **Clinical areas**
  - Clinician based data sources

- **External sources**
  - General Practitioners
  - Emergency Services
  - Allied health enterprises

Data Utilization

- Data collection management and organisation
- Knowledge discovery (population or cohorts)
- Understanding and comparison of cases
- Pre-processing and reshaping
- Statistical correlation and analysis
- Data aggregation and integration
- Identify events and trends
- Health awareness and promotion
Problems to Overcome for Data Linking

- Major practical impediments exist for data linking
  - Patient security and privacy restrictions
  - Diversity and independence of databases
  - Complexity of data formats
  - Sophistication of aggregation methods

- Existing solutions tend to adopt a heavy approach
  - Manual processing to achieve one-off linking
  - Data repositories or warehousing
  - Trusted third party units offering linking services
  - Full scale integration and interoperability of all systems

The HDI software tool
The Health Data Integration Project

- Aims to provide novel methods for linking multiple databases, by allowing data custodians to retain control of data
- Processes queries and reporting operations remotely at the data locations (in situ)
- Allows privacy and security restrictions to be met
- Enables audit trails of data access operations and users to be produced
- Fully software engineered application product has been produced at EHRC after about 20 person-years of effort

The HDI Solution
HDI: Data Custodian Control

- Data custodian retains control and security
  - No warehousing of data
  - All patient-identifying data is encrypted
- Databases are added to a HDI installation by the data custodian
  - The data custodian specifies who can use the data and how they can use it
- Metadata layer linked to industry standards to provide a common language and across repositories – increasing usability

HDI: Delineation of Responsibility

- HDI Domain concept
  - Provides demarcation of roles and responsibilities
    - Domain Administrator
    - Data Custodian
    - Project Administrator
    - Project member
  - Supports existing ethics committee approval process
HDI: Building a virtual data collection

Query on linked CRC surgical data and chemotherapy data to get a dataset of information on CRC patients, their current status and their chemotherapy treatments.

Query results show CRC surgical data and chemotherapy data for each de-identified patient.
HDI: Performing an Analysis

- Generate the survival chart, for example, Kaplan Meier.
- Analyse survival outcomes by stage.

HDI: Generating a report

Data from multiple databases can be used for reporting.

Eg… linked data from surgical databases, chemotherapy records and cancer registries.
Current Projects

Data Linking Projects

Some Current Clinical Applications of HDI
- Queensland Health: Queensland Oncology On Line
- Royal Melbourne Hospital: Colorectal Cancer
- Sydney South West Area Health Services: Colorectal Cancer
- Automated Cancer Staging: Lung Cancer
**Objective**
- Integration of CRC screening, surgery and family history information:
  - For research on sensitivity and specificity of the Faecal Occult Blood Test (FOBT)
  - For research on surgical outcomes including factors such as adjuvant therapy and co-morbidities

**Cohort**
- Surveillance program patient data over 25 years (approx 3000 records)
- Surgical data of approx 4000 records; Family history approx 4000 records

**Databases**
- 3 databases (CRC Screening Program (administrative and outcomes data), Surgical and Clinical CRC data, Family History data)

**Data Quality**
- Significant cleaning work required on databases to regularise recorded data and remove data entry, data format, or historical information change errors
Data cleaning effort and data analysis gave access to 25 years of surveillance information
- FOBT sensitivity published

Surgical Outcomes
- By factors such as smoking history, adjuvant therapy or diabetes status

### Surgical Outcomes

**FOBT sensitivity by colonoscopy findings in asymptomatic at-risk people**

<table>
<thead>
<tr>
<th>Colonoscopy Findings</th>
<th>FOBT Positive Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>64.5%</td>
</tr>
<tr>
<td>Adenoma</td>
<td>89.3%</td>
</tr>
<tr>
<td>Carcinoma</td>
<td>100%</td>
</tr>
</tbody>
</table>

- 32 patients had a preceding FOBT within 3 months (see Table)
- 5 cancers were detected, with 3 (60%) being FOBT positive
- 893 asymptomatic patients with planned colonoscopies

### Objective
- Demonstrate the ability to gather patient information across databases and across hospitals to provide summary information on quality and safety of patient care, adherence to clinical guidelines and comparison across hospitals
- To be able to provide this information to clinicians for access to their own data for case management AND for monitoring as above

### Cohort
- Approx 1000 ‘dummy’ patient records with CRC and corresponding administrative and treatment information

### Databases
- Different databases (Registry, Administrative, Surgical, Chemotherapy) for 3 different hospitals

### Data Quality
- Constructed data included deliberate errors in patient demographics and different formats of clinical information to demonstrate HDI’s ability in inexact matching and transformation of data entries to a selected standard
We have demonstrated that with HDI it is possible to report on indicators that require information to be linked within and across hospitals.
Staging of a cancer requires access to all available data:
- Radiology and histology text reports
- Information extracted from other forms of data, for example, radiological images

**Cancer Stage Interpretation System (CSIS)**
- Cancer staging is necessary to determine effective care for individual patients, as well as to design and evaluate health programmes at a population level
- Develop improved ways to access and analyse stored medical images and reports to better facilitate the staging of cancer patients
Future directions

Clinical to Genomic

- Integrating biomarker data sources and patient data can provide information on the efficacy, safety and toxicity of drugs
  - Advanced non small-cell lung cancer - Iressa effective in only 10% to 15% of patients. Scientists pinpointed mutations in a gene within some tumor cells that allows Iressa to work
- Micro Array experiments may be useful in finding disease biomarkers
  - Linking micro array results with known clinical outcomes will allow new biomarkers to be found
EHRs and Virtual Registry

- **Pre-population of Electronic Health Record**
  - Use data linking to gather as much information as possible (to required health record specification) about an individual to pre-populate their complete health record
  - Outcomes:
    - Reduced manual workload for initial set up of health record

- **Virtual registry**
  - Use data linking to gather registry data set information about patients from existing sources (rather than building brand new sources with paper based information submission) – and present back a view of the registry for administration and analysis
  - Outcomes:
    - Significantly reduced manual and paper based workload for clinicians and registry officers resulting in more timely registry information, and greater compliance with registry requirements if data capture is at source
Contemporary ICT Advances

- Web Services
  - Easy(ier) federation of data and services
- Ontologies
  - Relating concepts using semantic properties
  - Discipline based
    - Need domain expertise to define the ontology of data sources and fields
- The semantic web
  - Using ontologies in the web applications

Using SNOMED CT

- Use SNOMED CT for mapping of data to terms
  - Structured data sources
  - Natural language
    - Reports
    - Discussions
- To make this possible
  - Subsets for particular domains
  - Augmenting the domain specific subsets
  - Fast querying
SNOMED-CT Scope

- Clinical Terms
- Comprehensive, not specialty or domain
  - Human, veterinary, drugs, social, disease, observations, interventions and wellness
- ~400,000 concepts (fully specified names)
- ~1M descriptions (synonyms etc)
- ~1.4M relationships (900,000 defining)
- URU principles
  - Useable, Repeatable, Understandable

SNOMED CT Top-level Concepts

- 138875005 | SNOMED CT Concept
  - 123037004 | body structure |
  - 404684003 | clinical finding |
  - 243796009 | context dependent category |
  - 308816002 | environments and geographical locations |
  - 272379006 | event |
  - 106237007 | linkage concept |
  - 363787002 | observable entity |
  - 410007006 | organism |
  - 373873005 | pharmaceutical / biologic product |
  - 78621006 | physical force |
  - 260787004 | physical object |
  - T13688002 | procedure |
  - 362981000 | qualifier value |
  - 419691008 | record artifact |
  - 48176007 | social context |
  - 370115009 | special concept |
  - 123038009 | specimen |
  - 254291000 | staging and scales |
  - 105590001 | substance |
Expressions, equivalence and subsumption

- 85189001|acute appendicitis
  - subsumed by
    - 74400008|appendicitis|
    - 64572001|disease|:
      116676008|associated morphology|=23583003|inflamation|
      363698007|finding site|=66754008|appendix structure|
  - equivalent to
    - 74400008|appendicitis|:
      260908002|course|=53737009|acute|

Querying

- Simple queries can use subsumption/equivalence, for example
  - find all cases of 74400008|appendicitis| with a particular treatment or outcome
  - finding those also classified as
    - 85189001|acute appendicitis
    - or even as a 64572001|disease| with 23583003|inflamation| of the 66754008|appendix structure|
HDI Platform Technology

De-identified linked data for analysis

HDI integrates data

Custodial controlled data

HDI Hub

Statistical Packages
e.g. R, SPSS

Reporting Tools
e.g. Crystal Reports

Custom Applications

De-identified virtual linked data set

HDI Data Source

HDI Data Source

HDI Data Source