An Ontology-based System for Information Extraction, Reasoning and Cancer Registration from Pathology Reports

Giulio Napolitano, A Marshall, P Hamilton, C Fox, A Gavin

NCIN-UKACR, June 2012

Background

- Cancer Registry targets
  - 70% cancers staged
  - Timeliness
  - More electronic data captured

- Electronic sources increasingly available
  - MDTs
  - Radio & Chemo
  - Pathology
  - GPs
- Information Extraction (IE): a growing issue!

Project - Focus

- Surgical pathology
  - The most accurate source of information on a patient's cancer
  - Narrative or almost-narrative form
- Difficult to read by a machine
  - Requires visual inspection in almost all scenarios of information extraction
- ~1000 breast cancer reports in NI, in 2006
  - ~47,000 cases in UK

Semistructured reports

START_OF_REPORT
Secretary:- GN
CLINICAL DETAILS
Locally advanced Ca breast - Right T4, N1. Total mastectomy and axi
2 please. Tumour with inverted nipple.
PATHOLOGIST'S REPORT
1.
GROSS DESCRIPTION
TYPE OF SPECIMEN:Right total mastectomy.
SPECIMEN SIZE:
Dimensions:13 x 10 x 5 cm with an NBSE 13 x 7 cm.
Weight:309 grams
HISTOLOGY
HISTOLOGICAL TYPE:Infiltrating lobular carcinoma.
GRADE:II (3, 2, 2).
DCIS PRESENT:No.
SIZE OF INVASIVE COMPONENT:Estimated as at least 5 cm in diameter.
NEAREST MARGIN:Deep.
MARGINS: DISTANCE:
InvasiveIn Situ
Superficial:1 cm
Deep:<1 mm
Medial:2.5 cm
Lateral:4.5 cm
Superior:2.5 cm
Inferior:7 mm
LYMPHOVASCULAR INVOLVEMENT:Present, widespread.
AXILLARY LYMPH NODES:NumberInvolved
Level 1 (Part 2)1515
Level 2 (Part 3) 1212
Level 3 (Part 4) 77
EXTRANODAL DEPOSITS:Yes.

4

Unstructured reports

CLINICAL HISTORY SECRETARY: GM
Long-standing microcalcification. Felt to be benign. USS - R4. Core - blood++. Open biopsy. Palpable nodularity. Progressive calcification - core C5. Nodule UIQ separate from previous biopsy.

1. Right breast.
2. Level III nodes.

PATHOLOGIST'S REPORT :-
1. The specimen consists of a right mastectomy with attached axillary nodes. The specimen has been marked with 2 orientation beads. The red bead marks the axillary vein level and the purple bead marks the medial apex of axilla (level II). The mastetomy weighs 1033 g and measures 21 x 16 x 5.7 cm with a nipple-bearing skin ellipse measuring 20 x 13 cm. A horizontal healed scar measuring 6 cm in length is present in the region of the upper inner quadrant.

On sectioning the breast, a firm tumour nodule is identified within the upper inner quadrant measuring 2.3 x 2.0 x 1.5 cm. The tumour lies deep and superior to the healed scar. The tumour lies 13 mm away from the deep margin, 20 mm from the superior margin and 40 mm from the medial margin. The attached axillary fat from level I weighs 30 g - the attached axillary fat from level II weighs 25 g.

Histological examination of multiple representative sections shows the features of a grade III infiltrating ductal carcinoma (tubules 3, pleomorphism 3, mitoses 2). There is an area of high grade DCIS (comedo and cribriform) lying directly adjacent to the tumour but this does not increase the overall diameter

5

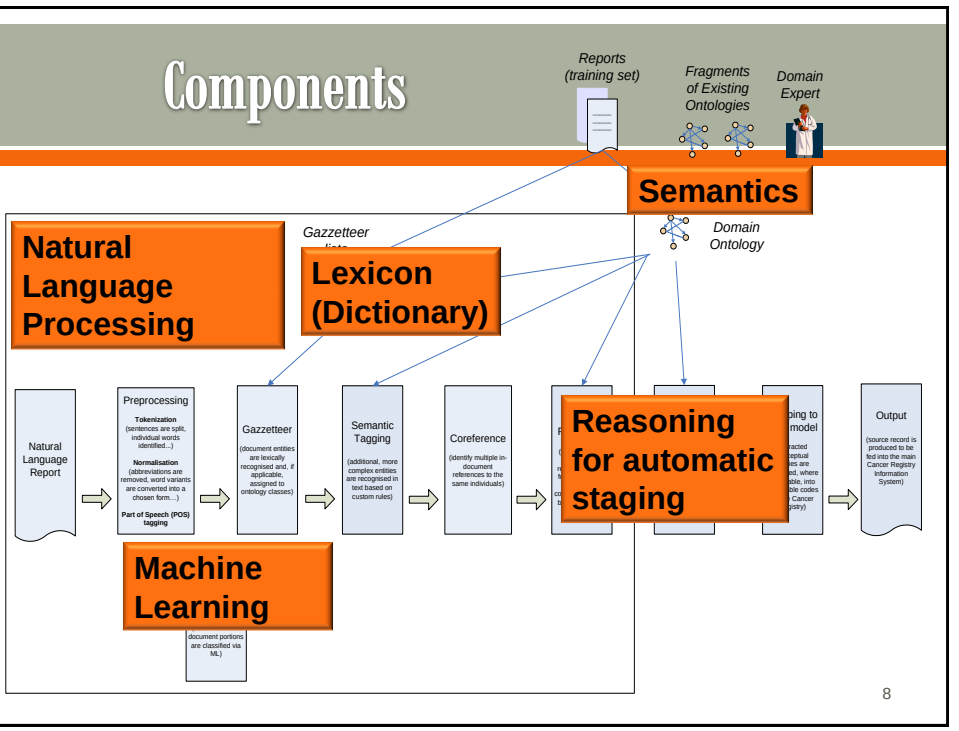
Project – Objectives and Tasks

- Objectives
  - Automated staging from pathology reports
  - Enhance IE, by means of an ontology-based approach
    - Additional fields: site, HER2 status…
  - Attempt to design a complete document-to-registration system
- Tasks
  - Produce a suitable ontology
  - Integrate with an Information Extraction (IE) system
  - Implement and test in the NI Cancer Registry
  - Integrate with the existing Cancer Registry system

Project - Benefits

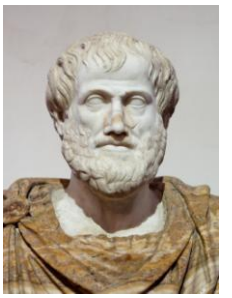
- Enhanced completeness and accuracy of cancer data
  - Stage at diagnosis, to improve
    - Assessment of treatment
    - Survival analysis
  - Diagnostic information, to improve
    - Incidence figures
    - Epidemiological research
- Provide further insight for general IE of fragmental biomedical text

Components

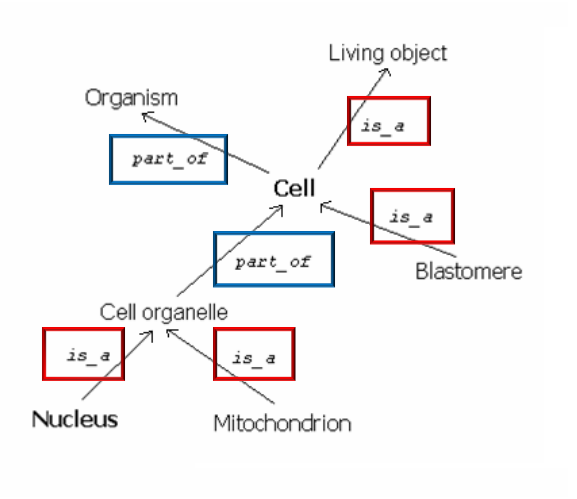


What is an ontology?

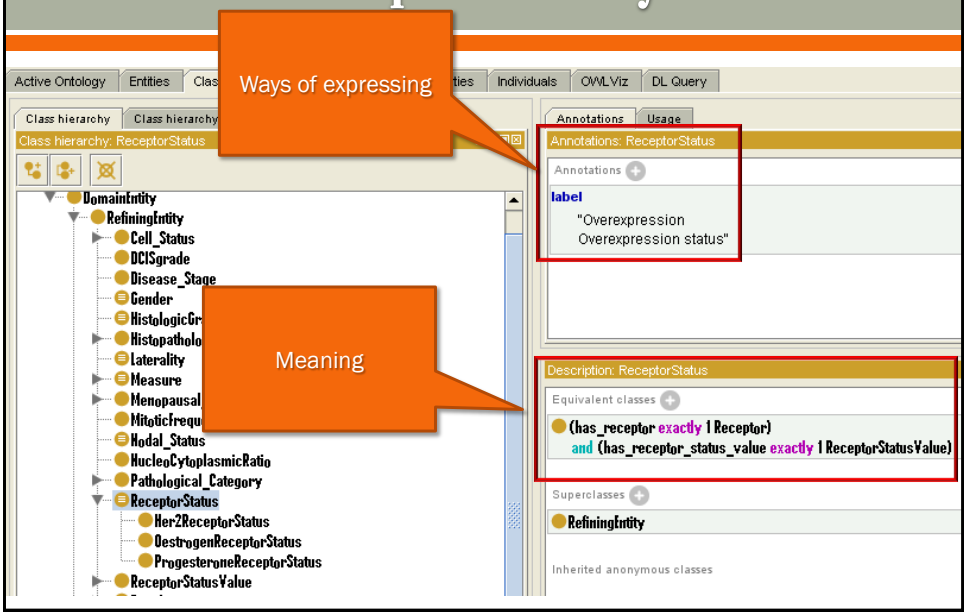
- Ontology
  - A philosophical concept
  - Going back to ancient Greece
- Ontology in Computer Science
  - Model of a domain of the world
  - Classes, individuals and properties
    - Hierarchical
- Machine-readable but also human-friendly



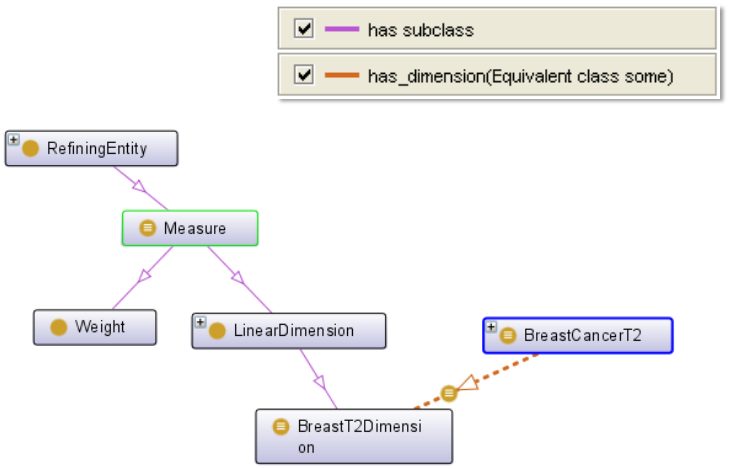
Sample ontology



Sample Hierarchy

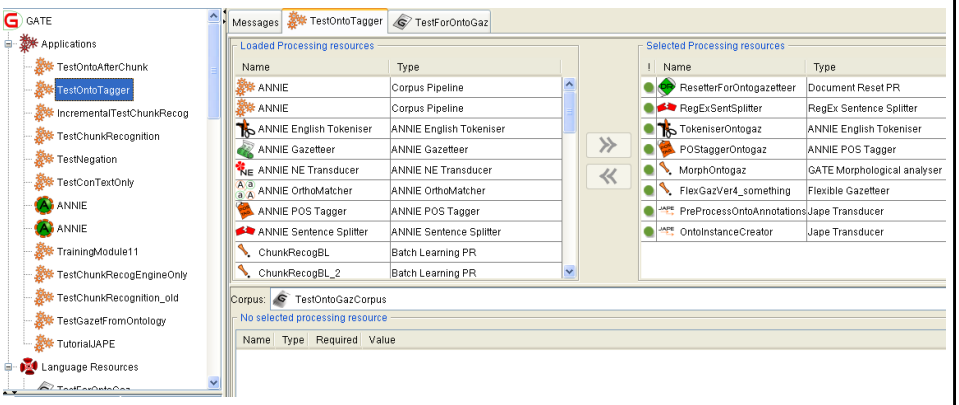


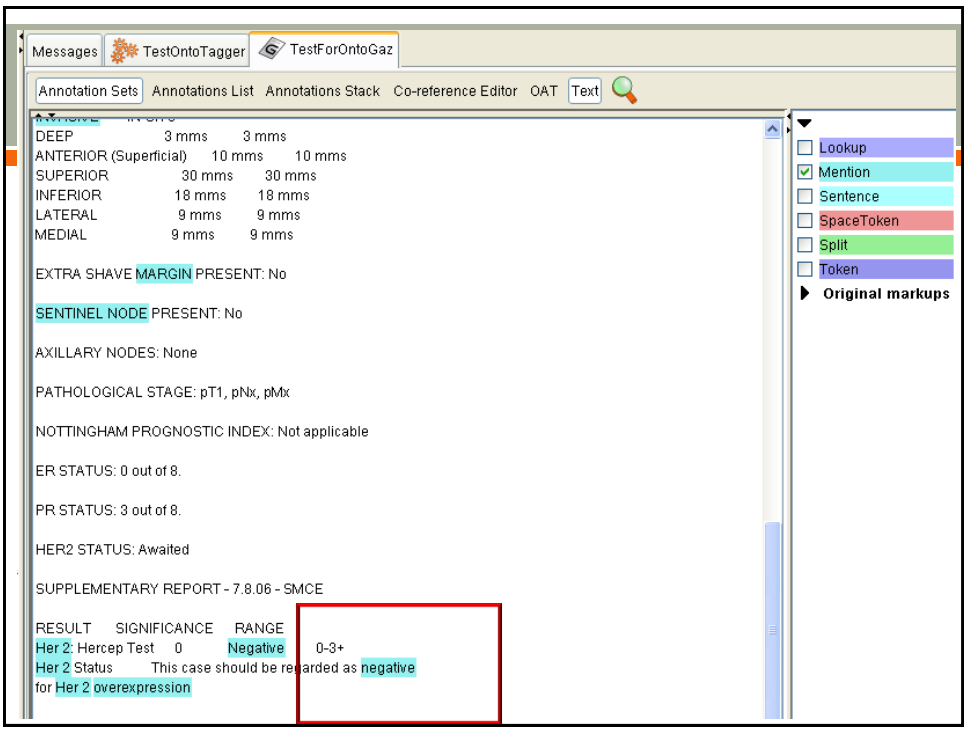
Defining Breast T categories



12

GATE


GATE
Applications
TestOntoAfterChunk
TestOntoTagger
IncrementalTestChunkRecog
TestChunkRecognition
TestNegation
TestConTextOnly
ANNIE
ANNIE
TrainingModule11
TestChunkRecogEngineOnly
TestChunkRecognition_old
TestGazetFromOntology
TutorialJAPE
Language Resources
Messages
TestOntoTagger
TestForOntoGaz
Loaded Processing resources
Name
Type
ANNIE
Corpus Pipeline
ANNIE
Corpus Pipeline
ANNIE English Tokeniser
ANNIE English Tokeniser
ANNIE Gazetteer
ANNIE Gazetteer
ANNIE NE Transducer
ANNIE NE Transducer
ANNIE OrthoMatcher
ANNIE OrthoMatcher
ANNIE POS Tagger
ANNIE POS Tagger
ANNIE Sentence Splitter
ANNIE Sentence Splitter
ChunkRecogBL
Batch Learning PR
ChunkRecogBL_2
Batch Learning PR
Selected Processing resources
Name
Type
ResetterForOntogazetteer
Document Reset PR
RegExSentSplitter
RegEx Sentence Splitter
TokeniserOntogaz
ANNIE English Tokeniser
POStaggerOntogaz
ANNIE POS Tagger
MorphOntogaz
GATE Morphological analyser
FlexGazVer4_something
Flexible Gazetteer
PreProcessOntoAnnotations
Jape Transducer
OntoInstanceCreator
Jape Transducer
Corpus:
TestOntoGazCorpus
No selected processing resource
Name
Type
Required
Value


Messages
TestOntoTagger
TestForOntoGaz
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
OAT
Text
DEEP 3 mms 3 mms
ANTERIOR (Superficial) 10 mms 10 mms
SUPERIOR 30 mms 30 mms
INFERIOR 18 mms 18 mms
LATERAL 9 mms 9 mms
MEDIAL 9 mms 9 mms
EXTRA SHAVE MARGIN PRESENT: No
SENTINEL NODE PRESENT: No
AXILLARY NODES: None
PATHOLOGICAL STAGE: pT1, pNx, pMx
NOTTINGHAM PROGNOSTIC INDEX: Not applicable
ER STATUS: 0 out of 8.
PR STATUS: 3 out of 8.
HER2 STATUS: Awaited
SUPPLEMENTARY REPORT - 7.8.06 - SMCE
RESULT SIGNIFICANCE RANGE
Her 2: Hercep Test 0 Negative 0-3+
Her 2 Status This case should be regarded as negative
for Her 2 overexpression
Lookup
Mention
Sentence
SpaceToken
Split
Token
Original markups

Evaluation protocol (training/test)

1. Is the system working as required?
2. Is the system producing the desired results?
3. Does the system work better than the existing procedure it will replace?
4. Is the system cost-effective?
5. What are the likely long-term impacts of the system?
   - E.g. recall/precision and specificity/sensitivity measures against performance of other tools and HUMANS
6. Further applications?
   1. MDT reports
   2. GP documents

Why do this?

- Interoperability
- Reusability
- Stability
- Versatility

16

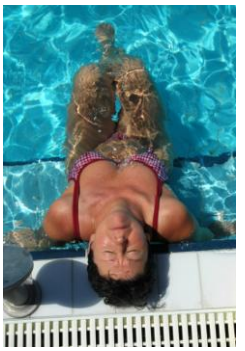
Don't forget to…

- Visit our Staging Tool stand
- View our related posters
  - Reasoning for staging (#92)
  - Staging Tool (#91)
- Follow us on twitter @NICanReg

17

Still awake?

- Questions…



18

An Ontology-based System for Information Extraction, Reasoning and Cancer Registration from Pathology Reports

Giulio Napolitano, A Marshall, P Hamilton, C Fox, A Gavin

NCIN-UKACR, June 2012