

The University of Manchester

MANCHESTER
1824

Knowledge Acquisition

COMP62342
Sean Bechhofer
University of Manchester
sean.bechhofer@manchester.ac.uk

The University of Manchester

MANCHESTER
1824

Knowledge Acquisition (KA)

- **Operational** definition
 - Given
 - a **source** of (declarative) knowledge
 - a **sink**
 - KA is the **transfer** of declarative statements from source to sink
 - we can generalise this to other sources, e.g., sensors
- We distinguish between KA and K **refinement**
 - i.e., **modification** of the statements in our sink
 - But this distinction is merely conceptual
 - Actual processes are messy
- Range of **automation**
 - Fully manual (what we're going to do!)
 - (Fully) automated
 - Possibly plus refinement
 - e.g., machine learning, text extraction

2

The University of Manchester

MANCHESTER
1824

From Knowing to Representation

- Source
 - A person, typically called the **domain expert** (DE, or “expert”)
 - domain, subject matter, universe of discourse, area,...
 - Key features
 - They **know a lot** about the domain (coverage)
 - They are **highly reliable** about the domain (accuracy)
 - They know how to **articulate** domain knowledge
 - Though not always in the way we want!
 - They have good **metaknowledge**
- Immediate Sink
 - A document encoded in **natural language** or semi-NL
- Ultimate Sink
 - A document encoded in a **formal/actionable KR** language
 - I.e., an OWL Ontology!
- This KA is often called Knowledge **Elicitation**

3

The University of Manchester

MANCHESTER
1824


Knowing to Representation

The diagram shows a flow from a source to an ultimate sink. On the left, a portrait of a man is shown with a thought bubble above it containing a cartoon of a woman and a child. The thought bubble is labeled 'Source'. An arrow points from the thought bubble to a light blue box containing the text 'Margaret Grace Rever is the mother of Robert David Bright'. Below this box, a downward arrow labeled 'Immediate Sink' points to a yellow box containing the text 'Robert_David_Bright_1965 hasMother Margaret_Grace_Rever_1934'. Below the yellow box, the text 'Ultimate Sink' is written.

4

The University of Manchester

MANCHESTER
1824



...there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns -- the ones we don't know we don't know.

5

The University of Manchester

MANCHESTER
1824

Eliciting Knowledge

- Proposal 1: Ask the expert nicely to write it all down
- Problems:
 1. They know too much
 2. Much of what they know is tacit
 - Perhaps can give it on demand, but not spontaneously
 - I.e., it's there but hard to access
 - They can't describe it (well)
 3. They know too little
 - E.g., application goals
 - Target representation constraints
 - E.g., the language
 - Their knowledge is incomplete
 - Though they maybe able to acquire or generate it
 4. Expense
 - Busy and valuable people
 - They get bored

6

The University of Manchester

MANCHESTER
1824

The Knowledge Engineer (KE)

- Key Role
 - Expertise in KA
 - E.g., **elicitation**
 - Knows the **target formalism**
 - Knows knowledge (and software) **development**
 - Tools, methodologies, requirements management, etc.
- Does **not necessarily know** the domain!
 - Though the KE may also be a DE
 - Most DEs are not KEs
 - Though they may be convertible
 - May be able to “become (enough of an) expert”
 - E.g., if autodidact or good learner with access to classes
- **Investment** in the representation itself

7

The University of Manchester

MANCHESTER
1824

Elicitation Technique Requirements

- Minimise DE's **time**
 - Assume DE **scarcity**
 - Capture **essential knowledge**
 - Including metaknowledge!
- Minimise DE's KE **training and effort**
 - Assume loads of **tacit knowledge**
 - Thus techniques must be able to capture it
- Support **multiple sources**
 - Multiple experts (get consensus?)
 - Experts might point to other sources (e.g., standard text)
- KEs must **understand enough**
 - So, the techniques have to allow for **KE domain learning**
 - KRs reasonably accessible to **non-experts**
- Always assume DE **not invested**
 - I.e., that you care more about the KR, much more

8

The University of Manchester

MANCHESTER
1824

Note on generalizability

- Many KA techniques are very specific
 - Specific to source (e.g., learning from relational databases)
 - Specific to targets (e.g., learning a schema)
- Elicitation techniques are generally flexible
 - Arbitrary sources and sinks
 - In both domain and form
 - NL intermediaries help
 - “Parameterisable” is perhaps more accurate

9

The University of Manchester

MANCHESTER
1824

Elicitation Techniques

- Two major families
 - Pre-representation
 - Post-(initial)representation
- Pre-representation
 - Starting point! Experts interact with a KE
 - Focused on “protocols”
 - A record of behavior
 - Protocol-generation
 - Protocol-analysis
- Post-representation (modelling)
 - Experts interact with a (proto)representation (& KE)
 - Testing and generating

10

The University of Manchester

MANCHESTER
1824

Pre-representation Techniques

- Protocol-generation
 - Often involves **video** or other recording
 - Interviews
 - **Structured** or **unstructured** (e.g., brainstorming)
 - Observational
 - **Reporting**
 - Self or shadowing
 - Any **non-interview observation**
- Protocol-analysis
 - Typically done with **transcripts** or notes
 - But direct video is fine
 - **Convert** protocols into protorepresentations
 - So, some modelling already!
- We can **treat many things as protocols**
 - E.g., Wikipedia articles, textbooks, papers, etc.

11

The University of Manchester

MANCHESTER
1824

Modelling Techniques

- (Often characterized by aspects of the target (OWL in our case))
- Being **picky**
 - Pedantic refinement
- **Sorting** techniques
 - are used for capturing the way people compare and order concepts, and can lead to the revelation of knowledge about classes, properties and priorities
- **Hierarchy-generation** techniques
 - such as laddering are used to build taxonomies or other hierarchical structures such as goal trees and decision networks.
- **Matrix-based** techniques
 - involve the construction of grids indicating such things as problems encountered against possible solutions.
- **Limited-information** and **constrained-processing** tasks
 - are techniques that either limit the time and/or information available to the expert when performing tasks. For instance, the twenty-questions technique provides an efficient way of accessing the key information in a domain in a prioritised order.

12

The University of Manchester

MANCHESTER
1824

Other Modelling Techniques

- Scenario descriptions
- Diagrams
- Problem solving
- Teaching
- Role Play
- Joint Observation
- Etc.

13

The University of Manchester

MANCHESTER
1824

Example: An Animals Taxonomy

- Task:
 - generate a **controlled vocab** for an index of a children's book
- Domain:
 - **Animals** including (think of these as CQ)
 - Where they live
 - What they eat
 - Carnivores, herbivores and omnivores
 - How dangerous they are
 - How big they are
 - A bit of basic anatomy
 - » legs, wings, fins? skin, feathers, fur?
 - ...
 - (read the book!)
- Representation aspects
 - Hierarchical list with priorities

14

The University of Manchester

MANCHESTER
1824

Protocol Analysis

- From interviews/**behaviour** to **analysable items**
 - Text! Text is good!
- From a text,
 - find **key** terms
 - **harmonise** them
 - capitalisation, pluralization (or not), orthography, etc.
- Keep **track** of
 - **Significance**
 - Core or peripheral terms
 - Illustrative? Defining?
 - **Situation**
 - Sentences or sections
- Output: List of Terms

15

The University of Manchester

MANCHESTER
1824

Animal taxonomy Term Generation!

Horse, Grass, Sheep, Goldfish, Trout, Wolf, Shark, Cow, Cat, Herring, Wheat, Bear, Dog, Tree

16

The University of Manchester

MANCHESTER
1824

Sort of Knowledge

- “Declarative” Knowledge about Terms (or Concepts)
 - Aka **Conceptual Knowledge**
- Initial steps
 - **Identify** the domain and requirements
 - **Collect** the terms
 - Gather together the terms that describe the objects in the domain.
 - Analyse relevant sources
 - Documents
 - Manuals
 - Web resources
 - Interviews with Expert
- We’ve **done that!**
- Now some **modelling**
 - Two techniques today!
 - Card sorting
 - 3 card trick

17

The University of Manchester

MANCHESTER
1824

Card Sorting!

- Card Sorting identifies **similarities**
 - A relatively informal procedure
 - Works best in small groups
- **Write down** each concept/idea on a card
 1. **Organise** them into piles
 2. **Identify** what the pile represents
 - New concepts! New card!
 3. **Link** the piles together
 4. **Record** the rationale and links
 5. **Reflect**
- **Repeat!**
 - Each time, note down the results of the sorting
 - Brainstorm different initial piles

18

MANCHESTER 1824

The University of Manchester

Sorted Animal Cards

The diagram illustrates the sorting of animal cards into three categories: Animal, Plant, and Fish. The 'Animal' category includes Horse, Sheep, Cat, Wolf, Bear, Cow, and Dog. The 'Plant' category includes Wheat, Grass, and Tree. The 'Fish' category includes Herring, Goldfish, Shark, and Trout.

MANCHESTER 1824

The University of Manchester

Try 2 Rounds

- Initial ideas
 - How we use them
 - Ecology
 - Anatomy
 - ...

20

The University of Manchester

MANCHESTER
1824

Generative

- For elicitation, **more** is (generally) better
 - Within limits
 - Brainstormy
- Is **critical** knowledge tacit?
 - We can't easily know in advance
- **Winnowing** is crucial
 - Sometimes we elicit things which should be discarded
 - And trigger the discarding of other things!
 - Better to know what we don't care to know!

21

The University of Manchester

MANCHESTER
1824

Knowledge Acquisition (KA)

- **Operational** definition
 - Given
 - a **source** of (propositional) knowledge
 - a **sink**
 - KA is the **transfer** of propositions from source to sink
- **Elicitation** (for terminological knowledge)
 - Initial **Capture**:
 - Source: People, "experts", "domain experts" (DE)
 - Sink: "**Protocol**" (record of behavior)
 - Term **Extraction**:
 - Source: Text (e.g., transcript, textbook, Wikipedia article)
 - Sink: **List of terms** (perhaps on cards)
 - Initial **Regimentation**:
 - Source: List of terms (on cards!)
 - Sink: **Proto-representation**
 - Hierarchy of categorized, harmonised terms (with notes!)

22

The University of Manchester

MANCHESTER
1824

Triadic Elicitation: The 3 card trick

- **Select** 3 cards at random
 - **Identify** which 2 cards are the most **similar**?
 - Write down **why** (a similarity)
 - As a new term!
 - Write down **why not** like 3rd (a difference)
 - Another new term!
- **Helps** to determine the **characteristics** of our classes
 - **Prompts** us into identifying differences & similarities
 - There will always be two that are “closer” together
 - Although **which** two cards that is may differ
 - From person to person
 - From perspective to perspective
 - From round to round

25

The University of Manchester

MANCHESTER
1824

Example

1. David Bright (1934)
2. Margaret Grace Reeve (1934)
3. Robert David Bright (1965)

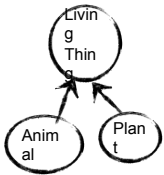
26

The University of Manchester

MANCHESTER
1824

20 Questions

- Like the **game!**
 - The KE **picks an object**/concept in the domain
 - The DE **tries to guess** it
 - and asks a series of **yes/no questions**
 - “Is it an animal?” “Is it a vegetable?” “Is it a mineral?”
- KE notes the **questions** and their **order**
 - Can help determine **key concepts**, properties, etc.
 - Animals, vegetables, and minerals!
 - Can help **structure** the domain
 - “Is it a living thing?”, “an animal?”, “a plant?”
- Note that the technique is not the game!
 - Goals are different!
 - We’re very interested in the questions, not the answers per se



27

The University of Manchester

MANCHESTER
1824

Key Goal: Laddering

- Terms **vary** in generality
 - Tree **vs.** Plant
 - Dog **vs.** Rover
- Each sort may be **implicit!**
 - Goal: **Flesh out** the generality **hierarchy**
 - Get more specific (if too general)
 - Get more general (if mostly specific)
- How?
 1. Take a group and ask **what they have in common**
 - During sorting or 3-card or directly
 2. Then **investigate relations** of new term
 - Siblings, missing children, and (eventually) parents (back to 1)

28

The University of Manchester

MANCHESTER
1824

So! The Task

- Capture
 - Look at the Menu
- Extract
 - List of terms; put them on cards!
- Organise
 - Hierarchy
- Encode
 - OWL in Protégé

33

The University of Manchester

MANCHESTER
1824

Coursework

- Take the KE done in class
 - Feel free to refine it further
- Encode it using Protege 4
 - Each category term becomes a class
 - Capture your hierarchy using subsumption/subclassing
- Submit your RDF/XML file
- Full description on Blackboard!

34