

Mapping concepts in social surveys: the experience of the Data Chronicles project

Karen Clarke

(with Judith Aldridge and Phil Edwards)

University of Manchester

The problem

- Wealth of social survey data available for secondary analysis
- Administrative statistics: an under-used resource?
- How do you find data sources unless you already know about them (ie, have some subject area expertise)
- How do you recognise the important conceptual differences between multiple data sources addressing the same topic?

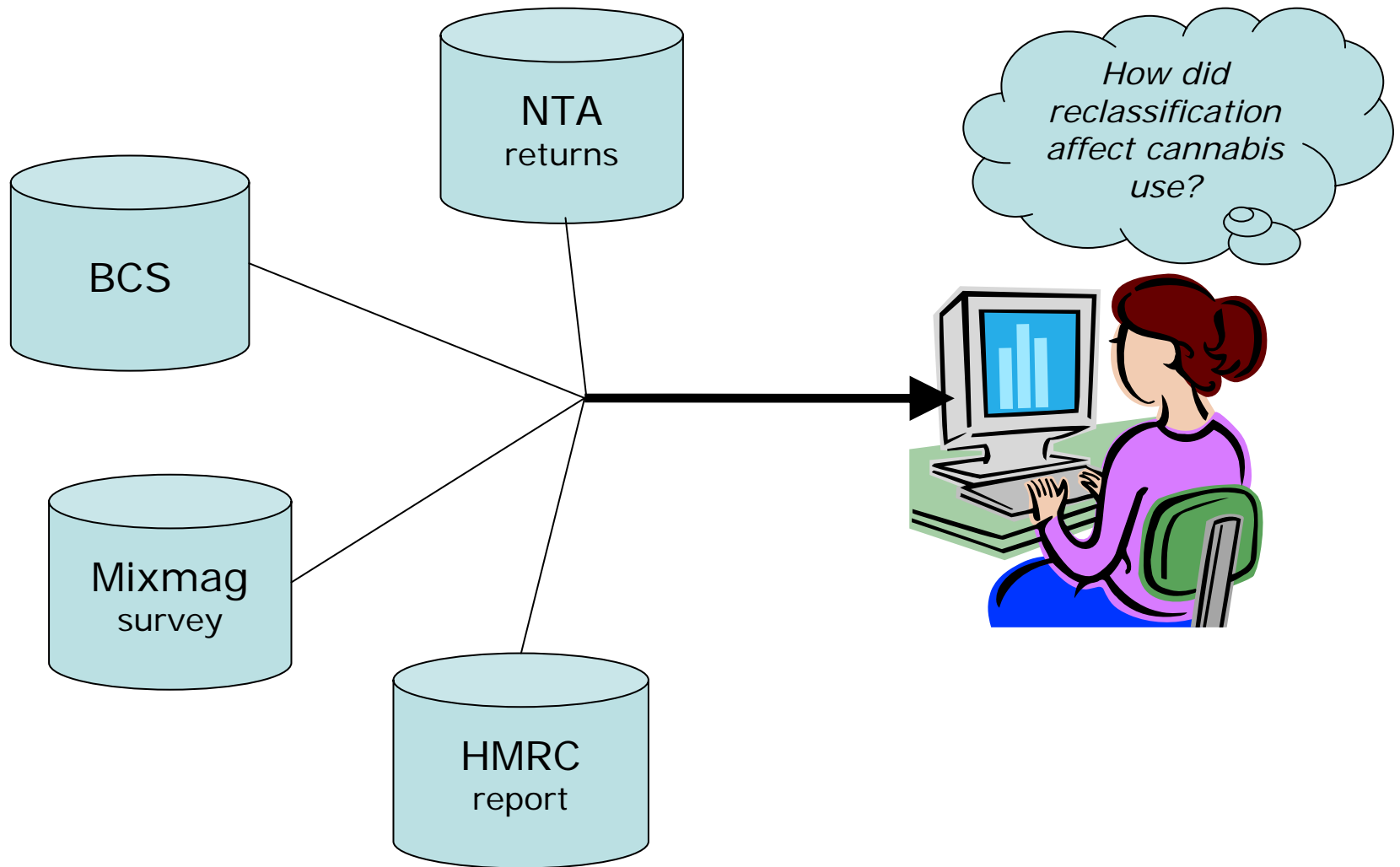
Vision for a Repository

- Build a searchable catalogue of metadata
- Deal with the problem that social science topics are approached from within different conceptual frameworks & different vocabularies
- Make accessible to users with different levels of prior knowledge

Aims of the pilot

- Create a limited repository in one field
- Facilitate the identification of data sources dealing with drugs, alcohol and tobacco
- Choice of this field:
 - multidisciplinary: health, crime, leisure/recreation

Background: putting it all together



Problems

- Different types of data source
 - Different levels of availability
- Survey data
 - Edited and published
 - Catalogued
 - Source instruments generally available
- Administrative statistics
 - Edited and published
 - **Not** catalogued
 - Sources **seldom** available

More problems

- **Imprecise concepts**
 - Definitions often vague
- **Contested concepts**
 - Multiple definitions of key terms
- **Mutable concepts**
 - Concepts change as society changes
- **But statistical sources still use them**
 - and there's no one right answer

The challenge

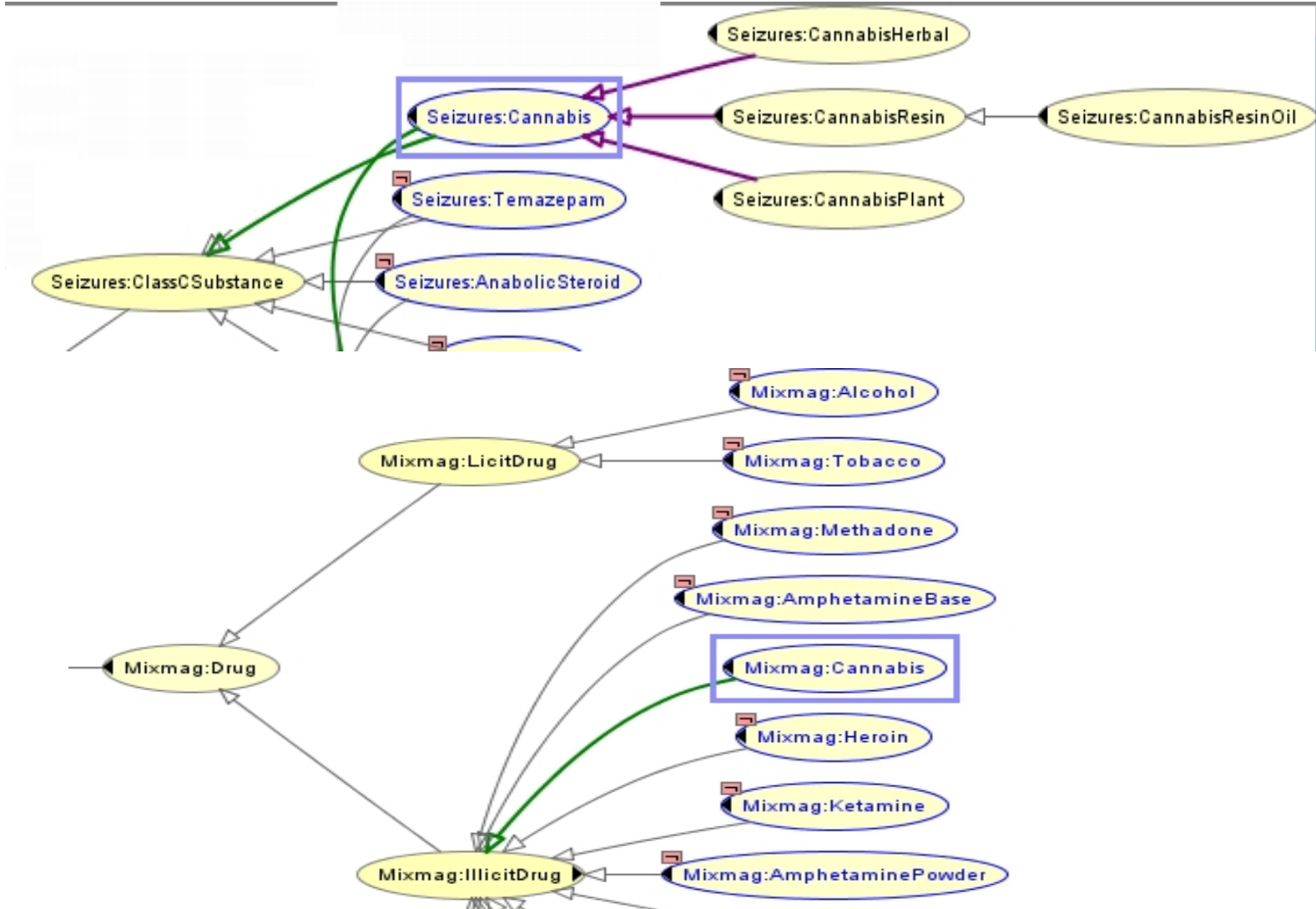
- We want to
 - Preserve the imprecision
 - Maintain multiple versions
 - Retain earlier generations
- We also want to
 - Make connections
 - Resolve confusion
 - Map the gaps

What the Respository does

- Maps the conceptual structures within a data source
- Links across data sources in a way that allows identification of common concepts and also differences

Two data sources: two hierarchies

HMRC seizures & Mixmag



Building the repository: creating a Babelfish

- Creating an ontology for each data source, mapping the structure of the 'in vivo' concepts using OWL protégé
- Constructing a 'lowest common denominator' linking concepts across data sources (the Babelfish)
 - A single conceptual hierarchy
 - Allows for cross-referencing between *in vivo* hierarchies

An example

- Multiple individual hierarchies
 - *Mixmag* says:
 - **Cannabis** is a type of **Drug**
 - *Seizures* says
 - **Cannabis Resin** is a type of **Class C Substance**
 - *NDTMS* says:
 - **Cannabis** is a type of **Problem substance**

An example (cont)

- Linking the multiple hierarchies
 - *Babelfish* says:
 - **Drug [Mixmag]** is a type of **Psychoactive**
 - **Class C Substance [Seizures]** is a type of **Psychoactive**
 - **Problem Substance [NDTMS]** is a type of **Psychoactive**
 - Moving up and down across hierarchies through Babelfish
 - Linking problem substances [NDTMS] to Class C substances [Seizures] to Drugs [Mixmag]

Remaining challenges

- ‘Higher order’ search terms
 - ‘Fuzzy’ search terms
 - Tagging as a solution?
- ‘Debate’ or information bubbles
 - ‘Information’ bubbles
- User modification
 - Desire paths (like in Amazon)
 - Other things

Demonstrator Repository

- You can see the demonstrator version at:
<<http://tinyurl.com/ttdpx>>