What are data archives?
Data archives are resource centres for analysts who use data for research and teaching. Their functions usually include:
- checking, validating and preparing data and accompanying user documentation
- cataloguing their technical and substantive properties for information and retrieval
- supplying them in an appropriate form to secondary users
- supporting users in using the data

What is social science data?
In the context of data archives, data means digital data. Data can consist of different types of qualitative or quantitative materials, for example numeric data, survey databases, interview transcripts, diaries, field notes, digitised materials, audio recordings, photographs and modelling scripts. Data collections are deposited with documentation and data collection materials.

How is data managed?
Data is managed by the Data Documentation Initiative (DDI), an XML-based descriptive metadata standard. The DDI is an international XML-based descriptive metadata standard for social science data used by most social science data archives in the world. A number of vocabularies control the content of the DDI fields.

The subject fields <subject> <keyword> are controlled in-house. The UK Data Archive maintains a vocabulary for its subject categories and thesaurus HASSET for its keywords.

Keeping vocabularies current
New concepts pose a challenge for maintaining the currency of subject categories for browsing. Data and its accompanying documentation will reflect new concepts found in the social science literature. Indicators or measurements of these concepts will also vary over time and from one research project to another. With concepts expressed in notation this problem is readily negotiated. UDC, with its online number building application, is quick and easy to use while at the same time its facility for building compound subjects provides sufficient detail in the code to allow for later disambiguation of subject categories, or the building of new ones, a relatively simple task.

Classification and indexing
After processing, the data is classified in a number of subject categories as best described by the research project, and HASSET key words are added to index the variables. The subject sub-categories are then mapped to a faceted list of subject categories.

Creating statistics
New subject categories can be tailored to meet special temporary needs such as:
- informing decisions about subject areas of the collection requiring development
- collecting material for workshops and other public events
- to gain an overview of the distribution of subjects within existing subject categories. For example in the subject category “Society and Culture” in the histogram we can see that “Leisure, Tourism and Sport” is emerging as a significant new category.

Conclusions: Social science data and UDC
UDC Online has resolved the main criticisms levelled at the UDC schema with respect to its application for indexing and retrieval in a digital environment.

Complex notation: Using UDC Online complexity in number building is no longer a problem and indeed is the feature of the schema that gives it the flexibility necessary to manage categories of complex social science concepts.

Dual approach to subject control: subject categories and HASSET thesaurus
Secondary data analysis involves the analysis of existing dataset(s). The project may require the location of suitable variables across a diverse selection of datasets. The combination of good explanatory tools such as Nessstar and powerful and accurate subject searches enable access to a rich pool of primary sources.

The dual approach provided by DDI fields <subject> <keyword> was designed to enable separate management of data ‘packages’ (identified by a study number) and the variables they contain. All studies are indexed to variable level.

Universal Decimal Classification: a pilot study
What proved particularly useful about the use of UDC for this project was that we could build numbers to include extra subject and use either a colon or a plus sign to indicate the relationship between subjects. When required we can retrieve these secondary subjects and allocate the dataset accordingly to the extra or new category, whilst still retaining all the required information in the full classification number attached to the study.

We included ‘form’ auxiliaries to explore the use of the notation to link data to resources that may be useful for training and outreach purposes e.g. case studies, illustrations/photos or bibliographical references. We also added auxiliaries of ‘place’ and ‘time’. The latter we thought would be useful for easy identification of historical data.

Revision: In recent years revision of notation in the social sciences is evident.

Legacy classification: For most repositories legacy classification of up to seven thousand items would not be feasible. However, while data collections are complex to index the subjects tend to be relatively generic; for example the Family Resources Survey. Thus the task of classification does not tend to pose the problems that may be encountered in the classification of a collection of documents of a similar size.