The LIGHTHILL RISK NETWORK

Data Catalogue Development

The purpose of this document is to set out the requirements of the Lighthill Risk Network's (LRN's) Data Catalogue due, in its initial form, at the launch of the "LRN Platform" dynamic site (testing in September; operational on October 31st). This proposal encompasses scope, approach, costs, and timings, together with any further features we hope to add.

Scope

One of the 'gold star' features of the web-based OSI, EPSRC and ESRC backed "LRN Platform" will be the Data Catalogue which will house Risk-related datasets (or hyperlinks to datasets) of relevance to the (re)insurance industry in categories of relevance to the industry. In most cases the data will exist as hyperlinks to the actual datasets downloadable from a remote location. However, in some instances the data will be located in the LRN Data Catalogue. In the latter case, the dataset will have been purchased in commonly used formats (to be defined via a user survey) by the LRN as an add-on service, accessible to paying subscribers only.

The types of data to be catalogued will be both UK and internationally sourced. Examples include meteorological data, hazard data, economic and social statistics (e.g. from the ONS), claims data, the list goes on. Geospatial data capable of linking into GoogleEarth and coded hazard events will be of particular interest as they enable scientific observations to be linked with insurance loss information. The datasets can be split up into low-hanging and high-hanging fruit in terms of their ease of acquisition. High hanging fruit might include an operational risks database for insurance (if there is one), employers liability claims in the UK, personal injury claims, fire losses in UK, catastrophe claims statistics (e.g. from the Property Claims Service – at a cost) etc.. Low hanging fruit may include meteorological and seismic data which are either publicly available or commercially purchasable, but relatively easily acquired once identified.

Potential issues that might arise during the course of this project include: how to best engage business in requesting the content and design of the Data Catalogue?; how much clarity do business have in their requirements?; how to identify the key questions and the right data to look at when providing an evaluation of the datasets?; and last but not least, the ownership/cost issues of the datasets.

The project will be divided up into three phases as detailed below.

Phase One (P1): industry data requirements

This will be an initial requirements phase that will involve polling the LRN's (re)insurance Core Members and other selected representative users (e.g. MAP and ING) to identify the key datasets and data types of value to their community. The approach will be of the 'blue skies' type such that polled individuals will be asked to provide their data 'wish-list', irrespective of how accessible or not this data may be. We wish to collect requests from actuarial teams, natural catastrophe teams, cat model developers, and underwriters. The survey will also ask for preferred data formats (e.g. netCDF, ascii, xls,…?) and data fields (e.g. lat, lon; T resolution,…?). The output of this phase will be a 'request list' from which to build the Data Catalogue.

Phase One (P2): build data-Catalogue w. data descriptions (due: end Sept. 2006)

This will involve the identification, acquisition and categorisation of key datasets on the user 'request list' from P1. The Data Catalogue will not only provide data and links to remote data, but will include data provenance and descriptions of the data to aid data manipulation (e.g. spatial and temporal limits and resolution, how the data has been treated or calculated – smoothed? trimmed? model reanalysis). The datasets/data-links will also be catalogued into categories for ease of navigation, and a 'search' feature will be provided to allow users to perform a keyword search through the descriptions of the data (which in turn point to the data). We may also consider a user star rating of the datasets with comments (c.f. Amazon) in one of the "LRN Platform's" collaborative workspaces.

Participating UK Research Councils (potentially BBSRC, EPSRC, ESRC, PPARC, NERC) will be asked to provide 'data' contact points at this stage, internally and within their data centres/institutes. The Data Catalogue user 'request' list will be cross-checked with records of datasets arising from Research Council funded research (housed within Research Council data centres and institutes, or held by funded researchers within universities). Proposals will potentially be agreed with ESRC, NERC and PPARC data centres, whereby the datasets are 'piloted' to evaluate their usefulness to (re)insurance users prior to incorporation into the Data Catalogue (for commercial purchase). The 'piloting' of UK research data will constitute part of Phase Three (P3).

Phase Three (P3): informed data evaluation (ongoing, initial release on Oct. 31st)

The data will be evaluated by a team of representative business recipients (e.g. an underwriter, an actuary, a loss modeller and a cat model developer; or ideally a working party to represent each) to assess the source, quality, format, consistency and usefulness of each dataset, with a view to making it usable by the industry user. The data evaluated by each representative user/working party will reflect the requests for data made by the matching user category in P1. In some cases the dataset evaluation note will be the final product. In the ideal case, standards will be applied to treat the data (if necessary) in order to produce a usable dataset (in terms of format, resolution and fields).

This phase will take place between September and October.

Subject to our discussions with the Research Councils (as outlined in P2), there may be an ongoing project to pull together 'useful' data from UK research with a joint project team. This work would take place between September and December (say).

Technology

The Data Catalogue will sit inside the front-end of the OSI, EPSRC and ESRC backed collaborative "LRN Platform" (Vignette based). The Data Catalogue links to remote datasets is straightforward html, and datasets (where acquired) can be uploaded into the front-end file database (allowing them to be spidered). Permissions to access the Data Catalogue can be set, as can permissions to download datasets. A transaction portlet will be part of the 'Version 7 July release' (Roy Williamson to provide details) that could allow members to purchase datasets. Similarly, transaction software could be integrated into the "LRN Platform".

Project Approach Management

Celine Herweijer will act as Project Manager. We hope to involve Jeff Park (Park Associates) and James Orr (Quantitative Finance Network) as advisors throughout this project.

Phase One (P1): 'requirements'

The Project Manager will coordinate the polling of the Core Member companies and other representative business users. The dataset 'request lists' will then be collated and categorised. On the basis of what we find an updated Project Definition will be produced.

Phase Two (P2)

A core project team will be put together consisting of:

1. Database Co-ordinator - to assist with identifying and locating datasets (and links to remote datasets) on user 'request list', and indexing them into pre-identified categories. Tasks include web-searching, liaising with centres/institutions for data; liaising with person who requested the data on 'usefulness', and data/information

entry. Candidate – Lloyd's to suggest? If not, other core member companies to suggest?

2. Database Technologist – to design Data Catalogue and implement into the "LRN Platform" (with the help of Roy Williamson and Natalie Compton). Candidate – Lloyd's to suggest?
3. Peter Taylor – reviewer

Following an introductory meeting on Aug. 23rd, the P2 core project team will meet 1Xweek to discuss progress, with sign-off in late-September.

Phase Three (P3)

A core project team will be put together consisting of the Phase Two team plus:

1. Representative underwriter/loss modeller - Jakir Patel (Faraday)
2. Representative actuary -?– Working Party of Actuaries?
3. Representative cat model user -? suggested - Jane Toothill and team (Guy Carpenter)?
4. Representative cat model developer -? suggestions ?

Working parties will be set up with a private workspace on the "LRN Platform" to aid collaboration and upload progress made.

It is anticipated that the representative recipients will attend one initial meeting in early September with the Phase Two core project team, plus collaborative workspace discussion in the interim with sign-off in mid-late Oct. in time for the operational release of the "LRN Platform". The expected time commitment for each representative recipient is 5-10 working days. Where possible, a 'representative recipient' will be represented by a working panel such that time commitments of each individual are considerably lowered (e.g. 1hr/week over Sept. and Oct.).

Costs and other resources

Aside from the purchasing of datasets (which will be made available for a fee), this project should be at NO COST to the LRN. Phase One team members in addition to the LRN staff we hope will be assigned by Paul Nunn (Lloyd's), whilst the Phase Two 'representative recipients' will be free. In addition, we will aim to engage Research Council staff to assist with locating and acquiring data held by their data centres, institutes or funded academics. In the longer term, we will investigate the possibility of obtaining Research Council funding to assist with staffing and development/resources costs of the project (e.g. to fund an LRN 'Data Manager''). Roy Williamson will be consulted on occasion to with respect to how the "LRN Platform" could support the Data Catalogue (e.g. search features).

Timescales

Whilst the collection of data will be ongoing, the major deliverables and check-points are as follows:

Phase One (P1)

- CH to email Core Member companies and other representative users to ask them to circulate a request for datasets within their respective organisations (start: Aug. 7th, due Aug. 21st)
- CH to send Data Catalogue project proposals to Research Councils (PPARC, BBSRC, EPSRC, NERC and ESRC), and suggest a follow-up meeting with Research Council 'data contact points' during P2.

- CH to collate dataset requests from Core Members and categorise (PT to assist). (due: Aug. 25th)

Phase Two (P2)

- Aug. 25th: 1st Meeting of the P2 Core Project Team
- Thereafter: weekly project team progress meetings.
- Meet with Research Council 'data contact points' to discuss involvement in Data Catalogue project and match user 'request list' with datasets funded by the Research Councils.
- End date: mid September (Data Catalogue incl. data descriptions in "LRN Platform")

Phase Three (P3)

- CH to make enquiries about availability of potential 'representative recipients' (first meeting scheduled of P3 project team in early Sept.; phase to run in September – mid-late October)
- **Initial release: mid October**