

The Digital Report Generating System

Accessing the geological textbase

by Tim Duffy, *Edinburgh*

Knowledge is at the core of everything the BGS does. That knowledge resides firstly, in BGS staff with their broad range of geoscientific expertise. Secondly, a great deal of information that contributes to the BGS knowledge base is in digital databases. Thirdly, much condensed and interpreted knowledge is conveyed visually in BGS geological maps, whether published on paper or in digital form. However, a fourth repository containing a great deal of knowledge is in the geoscientific texts that the BGS has produced since 1835 – and there are a large number of these. But how to get at this text-based knowledge easily?

The modern geoscientific researcher increasingly expects all information sources to be available in digital form – increasingly over the medium of the World Wide Web (WWW). A prototype system has been built on the BGS Intranet for making text-based knowledge more readily accessible. A series of descriptive geology texts: geological memoirs, technical reports and regional guides concerning an urban area which generates many queries – the Greater Glasgow area – have been converted into digital form. Rather than simply digitally scanning the text and packaging them up in digital formats for ease of modern distribution – as is increasingly being done with recently authored BGS texts – the documents were converted from word processor files or scanned images into pure text. They were then marked up with ‘tags’, inserted into the text, that describe the structure of the knowledge in the

document rather than the presentation of that knowledge on the page. The document structure is described using Standard Generalised Markup Language (SGML an ISO standard) or eXtensible Markup Language (XML, a recommendation of the World Wide Web Consortium). This allows a computer search engine to search within documents and return that specific piece of text being sought with any associated plates, figures, tables or references. This contrasts with a more conventional WWW search in which the researcher receives a set of documents that might possibly contain the knowledge being sought. The tags dividing the text into topics or ‘text objects’ define what the knowledge categories are. In the case of the prototype, all documents are tagged with codes that describe the chronos-

trigraphical and lithostratigraphical entities being described in the memoirs, as these are the basic units in which most readers of this particular type of document are interested. The marked-up documents form the text database or textbase which can now be searched.

A considerable added benefit of this tagging system is that the geospatial links (in this case the lithostratigraphical codes) between geological features being textually described and the BGS’s digital maps can be exploited so that the traditional first way of looking for such knowledge – reading the published map – can be used as a starting point for finding the related text. By interactively searching a web-based geological map, the researcher can use the prototype to define a specific area; the system will not only return those parts of texts that fit the topics chosen, but also only select text from those documents that cover the area of interest. Returned text and related figures can be brought into a word processor to create a customised report.

The tagging system used defines how the knowledge encapsulated within each document is categorised. Different tagging systems can be defined for different types of geoscientific document, forming the basis of a standard geoML tagging language for all geoscientific textual knowledge repositories.

For further details contact Tim Duffy on:

Tel: 0131 667 1000
E-mail: trd@bgs.ac.uk



Screenshot of the DRGS prototype implemented for the Glasgow area.

Geological information BGS © NERC. Topography based on OS data with the permission of Ordnance Survey on behalf of The Controller of Her Majesty's Stationery Office © Crown Copyright. All rights reserved. Unauthorised reproduction infringes Crown copyright and may lead to prosecution and civil proceedings. Licence number GD272191/2000.