

Data Harvesting

On-time, accurate and easy delivery of data

What is data harvesting?

The concept of data harvesting can be quite daunting to most users. It provides yet another theory to learn, more industry terms (round-tripping) to follow and, undoubtedly, there will be more acronyms to study. But the concept itself is really very simple.

In effect, data harvesting is the gathering of data from numerous disparate databases into a single database from which it can be re-published in a unified manner. Since the easiest way to read and publish data is in XML / GML format, the term also incorporates the concept of "schema translating" this data into XML / GML formats and delivering this data according to a particular schema.

Why is data harvesting important?

More and more, we are under pressure to make the information we hold in our databases available to centralised sources or "hub databases". The "hub" in turn collates the data from multiple suppliers and presents it in a unified manner – easily understandable to the outside world – often against a geographic background.

Whereas this used to be achieved through time-consuming paper trails and documents, the continued adoption of the internet as a source of information means that every one now expects up-to-date intelligence reporting all the time at the click of a button.

To ensure that the Hub Database is constantly up-to-date, the suppliers need a quick and easy mode of making their data available to the hub. Data harvesting has gone digital.

Local and Hub Database Relationships

To understand the concept of data harvesting it is important to reiterate the relationship between the Local Databases and the Hub Database.

The Local Database holds the information as collected from the field and input according to the business process of a particular organisation (organisations store similar data in different databases and often use different database models). The data then needs to be fed to the hub.

The hub is responsible for collecting – or harvesting - the data from multiple Local Databases, loading that data into a centralised database and then hosting the data and making it available to external users.

The Duty of the Local Database

In order to harvest data to a unified hub a common model for the exchange of data needs to be agreed so that the hub receives data in a consistent form from all the local databases. With an agreed common model, each Local Database then needs to be mapped to this model in order for it to be harvested by the hub. Defining a common model is ideally suited to XML Schema, which in turn makes XML / GML the ideal format for harvesting. Once the mapping to common model has occurred, the easiest way to respond to harvesting demands from the Hub Database is to make the data available to the internet (in some cases, intranet) via a Web Feature Server.

As such, the supply is achieved via these three easy steps:

- Step 1) Understanding of the common model
- Step 2) Mapping to the common model
- Step 3) Publishing the data via a Web Feature Server

The activities of the Local Database have been highlighted in pink and yellow (representing two independent organisations) in the below diagram.

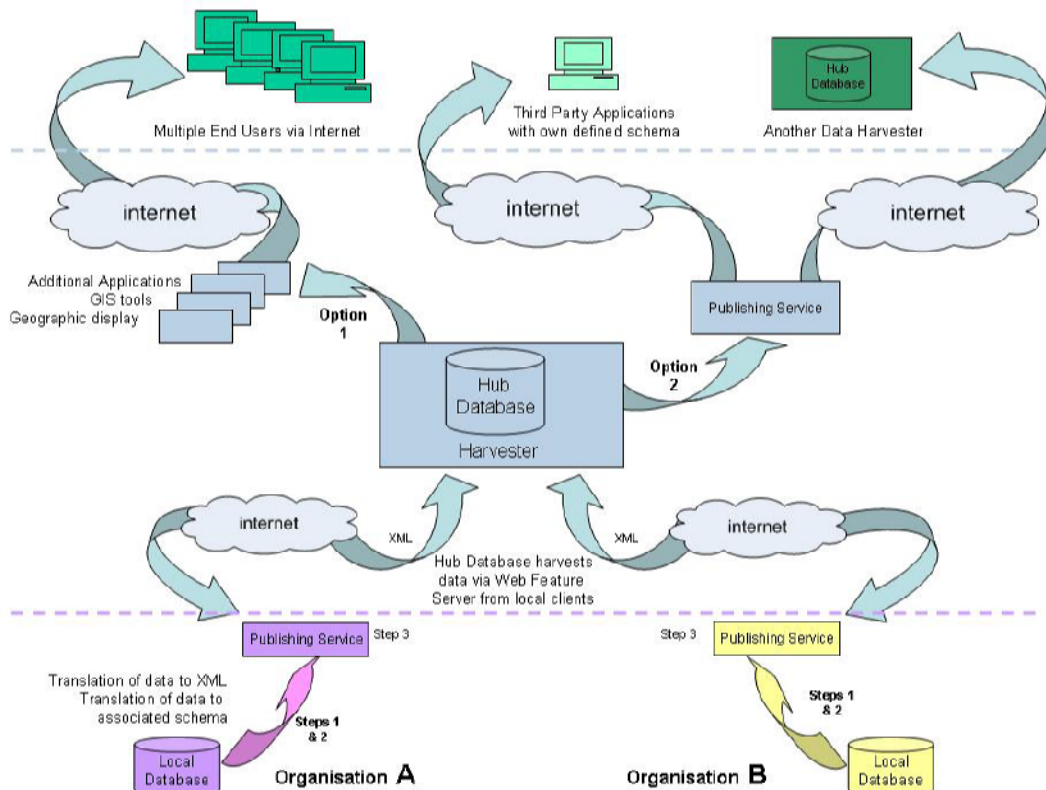


Figure 1: Details of a Data Harvesting Environment

Steps 1 and 2 take place locally within the respective organisation. An additional advantage of this is that these are one-off actions. Once the schema translation has taken place and a connection has been established via the Web Feature Server, the local database need only be updated and saved. The hub “harvests” all the suppliers periodically to gather any updates¹.

The Duty of the Hub Database

Per above, the first duty of the Hub Database is to periodically harvest the data from all the Local Databases, collate the data and load it into the Hub Database. Since mapping to the common model has already taken place at the Local Database, no further action need take place here.

Once hosted, the prime responsibility of the Hub Database is to make the collated data available to others. It can do this via either (or both) of the following options:

- Option 1) The harvested data can be collated and integrated with GIS applications to display the Information in an interactive and functional environment. Often, this application geospatially represents and displays the data. In a typical environment, this would then be hosted as a service for multiple end user access via the internet.
- Option 2) The Hub Database may be used to supply the collated data to various third party applications. The data has already been provided according to the model required by the hub, but it may have to be re-published according to one or more alternative models to meet the needs of different applications.

Alternatively, the Hub Database might be required to supply the collated data to another Hub Database. And so the cycle continues: the Hub Database becomes, in itself, a Local Database in a hierarchy of databases. A regional hub, for example, could supply data to a national hub.

Conclusion

By harvesting data into a central hub, end users of the data are provided with a unified view of the data even though it has been created in a disparate group of local databases. The local databases are responsible for providing data to the hub, but only the hub is responsible for providing data to the end users. This means that only the hub needs the scalability and high availability that may be required to support those users.

By translating the data into a common model for exchange, the local databases can continue to operate using their own local model. This allows existing databases to be brought into the network without modification and therefore without disruption to existing business activity.

Contact Information

Snowflake Software
T: +44 (023) 8023 8232
E: info@snowflakesoftware.co.uk
U: www.snowflakesoftware.co.uk

¹ An alternative option is to set up the Hub Server to receive data periodically from the suppliers. In this case, each local organisation is responsible for both the upkeep of the database and “serving” the updates to the hub on completion.