# Text Archive

Reuters has two separate text archives which can be made available to customers:

1. Reuters NewsML-G2 archive – this is for content since 2016 and is stored in NewsML-G2
2. Newsroom archive for content up-to October 2017 – this is in a format known as NRD XML (Novus Ready DTD compliant).

While both formats are XML based the second format is significantly different with Novus having a focus on text and search/retrieval so closer to an extended NITF format than the multimedia approach of NewsML-G2. The remainder of this document concerns delivery of NRD XML archive text.

## Delivery

When an archive text sale is made development will produce a customer specific zip-archive package of the required content that can be downloaded from AWS-S3. The link when unpacked will be a single text file representing all documents in the newsroom extraction. This will be a large (1GB+) file.

## Newsroom Archive Structure

The top-level item is an **n-archive-reponse** with a collection name and time. Within this are then **n-document** items for each text item in the archive each **n-document** having a unique **guid**.

Within the **n-document** there are two nested levels **n-metadata** and **n-docbody** the latter is where the news relevant content (and metadata) is located.

## Example Fields

Within the **n-docbody** common fields would be:

Headline: `n-document/n-docbody/document/title-info/title/`**`sort-title`**

Body text: `n-document/n-docbody/document/content/`**`text`**

(both <text type="lead-para"> and <text type="body"> will need to be retrieved

Within the supplier-info block there are various nitf-meta-content fields that will immediately look more familiar to those used to NewsML-G2 news content with fields like priority, transmitID, sent, etc.

The recommended date/time fields (original and latest publication) being:

`n-document/n-docbody/document/supplier-info/nitf-meta-content name="FirstCreated"` and `/nitf-meta-content name="VersionCreated"`. Note however in some older content the `nitf-meta-content` section is not present and the only publication date available is `n-document/n-docbody/pub-info/pub-date/sort-pub-date`. This omits the timestamp and is a UTC/GMT based publication date.

## Category Fields

Category information is maintained in language specific indexing elements, containing all keywords and subject terms in a specific language. Separate indexing elements are also maintained for journalist entered terms versus machine determined terms:

`n-document/n-docbody/document/indexing iso-language="en"`

and

`n-document/n-docbody/document/indexing iso-language="en" value-add="machine-generated"`

Within both sections there is normally an RCS block within a **classification-terms** element that gives category information in the RCS format: `http://s3.amazonaws.com/tr-liaison-documents/public/Reuters_News_Topics_External.xls` This is probably the best tag set to identify the story in terms of business sector, geography, organization, etc in order to produce a "training set" of data.