The system of data-collection on forms for automatic data-processing in archives

by

Enrica ORMANNI
Archivi dello Stato - ROMA

FOREWORD

Whenever an organic whole of homogeneous media is formed, the need arises of keeping it in such an order as to allow the retrieval and utilization of the units forming it; in other words, the need arises of carrying out, in respect of this complex, an information management that presupposes the study of the information sources from the viewpoint of their structure and typology.

A characteristic feature of the archival information science is that it is conditioned upon the structure of the set with which it deals.

An information science of a new type, that is to say, released from the conditioning resulting from the historical structure of files during the collecting procedure, is however applicable through automatic methodologies, conceived for facilitating a number of classes of searches that might be otherwise carried out only through the manual scanning of documentary texts.

The classes of searches allowed are dependent upon the choice of the data processing level; a choice that operationally turns into the study of the data-collecting methods and criteria, and that takes place mostly at the very stage of data-collecting.

DETERMINATION OF DATA PROCESSING LEVELS

The factors - interdependent - that determine the data processing levels are : language, extent of data-collecting, extent of information kinds and the retrieval aids that are to be made ready.

1.- language

This is the more determinant factor in respect of the A.D.P. level. The utilization of an artificial language, i.e., of a predetermined or open vocabulary of descriptors, certainly brings about an ADP level which hardly goes in depth, for it excludes an information retrieval on the textual elements. In fact, the use of descriptors is aimed at retrieving relevant information sources, while the appraisal of those containing pertinent information is referred to a non-automatic information retrieval.

The choice of operating at a natural language level(1) does instead allow to carry out an information retrieval of sources, also for the purpose of an appraisal of the pertaining information contained therein. This appraisal will be more or less pertinent depending upon the chosen extent of data collection.

2.- Extent of data collecting

Also this factor contributes to a remarkable extent toward the deepening of the data processing level.

Whenever an artificial language is used, the descriptor classes may be limited to the structural elements alone, namely, to those of them that refer to the historical structure of the documentary set; they may be extended to cover also the essential content of the information source, i.e., to the object of the activity that has led to the formation of one single archival entity(2); that may be pushed as far as to the indexing of some classes of information, relating to the essential content of the source in an indirect way.

Whenever a natural language is adopted, namely, whenever key-words are taken from the original texts, the deepening of the data processing depends only upon the actual quantitative extent of the classes of key-words that one decides to collect (an extent that may even include all the words of the documentary texts), but also upon the number of the selected leads of arrangement(3), i.e., upon the extent of the quantitative aids made available to the search.

3.- Extent of information kinds

The extent of information kinds is dependent upon the chosen source of data-collecting and upon its level in respect of the original source. If the conservation and retrieval aids manually processed at the time or after the formation of the documentary set are examined, we may have increasingly large levels, depending on wether one works on inventories, on analytical indexes, catalogues, calendars or abstracts. The classe. of information they contain are more or less comprehensive, according to the objectives for the attainment of which the information management was brought into being, but they are at any rate limited if compared with those that the original source may offer.

These first sources of data collecting (secondary in respect of the original ones) are in fact drawn up in artificial language; this brings about a non-deepened data processing level, even when the source used contains a large number of information items, as in the case of calendars. Moreover, the language used in these aids, at different times and with different purposes, is not standardized, and so no certainty can be adquired as to the semantic value of which the descriptors are the carrier, either at the time the data are collected or at the time in which the user performs, on the basis of them, appraisals as to the relevance of the original sources to which such aids refer. For these reasons, a treatment based on secondary sources is adopted only in special cases.

As a rule, the sources of data collecting adopted are the primary ones; as concerns archival groups, such sources may be the archival entity the single document, the single information item. The choice of the three sources of data collection conditions in part the language to be adopted, which, as far as the archival entity is concerned, is prevalently artificial, namely, a language of descriptors (alphanumeric codes, words, numerical data), either formatted or not; moreover, the extent of information kinds is limited by the very fact that they must at any rate comply with the essential content of the archival entity.

If the single document is instead taken as a source of data collection, the data processing level may be deepened using the natural language, that is to say, taking key-words from the text and extending the data collecting to cover all the classes of homogeneous data provided by the documentary set. All this turns into as many classes of information data.

Whenever one works at the level of single information items, pertaining to one or more principal objects of the documentary set (as in the case of the single land parcel in the field survey fonds), all the classes of information data contained in the sources are necessarily included.

4.- Retrieval aids

The retrieval aids may be processed on media man can readily read (such as printed outputs, micro-films, etc.), and may occur in the form of lists (either kwic or kwoc) or in the form of indexes, namely, of a group of words connected with one another by relations of various kinds and structured according to leads of arrangement. These aids, of an easy diffusion and reading, are, however, strongly oriented towards given classes of searches, unless complementary or introductory to other retrieval aids, capable of allowing a free search by means of terminals. In substance, they are all the more valid when they were conceived not merely as final product but also as intermediate finding aids.

The retrieval aid that involves more in-depth treatment levels is the one allowing an interactive access (either in batch or on line) to the data bank, and allowing the use by the user of any type of data utilization and the personal processing of aids particularly oriented towards his own searches. This calls for the arrangement of normative thesauri (either of descriptors or of key-words)(4) or of other aids affording the opportunity of using the data bank in the most complete way.

The arrangement of further aids, within the system (that is, managed by the program), may further deepen the data processing levels. The aids within the system may be oriented towards the automatic lemmatization, to the connection of graphic variants, to the processing of finding aids by the user itself.

DATA COLLECTING

That of data collection, in its moment of analysis and study of sources - introductory to the carrying-out of the collecting operations - is the phase in which the data processing levels, and hence the type of search to be performed and the processable aids, are determined. The aids that can be arranged for the interactive access are totally determined at this stage. The intermediate finding aids are instead determined only for what concerns the choice of the classes of data that will be a part of them, but not in the structure. The latter will result from analyses performed *a posteriori* on the data bank (by means of automatic procedures) in the light of the data emerging from the analyses and from a more in-depth knowledge of sources, acquired during the processing of the documentary set as a whole.

The moment of data collecting is the most professional one and, therefore, the general methodologies, that may be applied in part also to other documentary sets than archival fonds receive here applications typical of archival information management.

The conventional information procedures in respect of archives consist of two basic operations directed at the conservation and at the utilization of the archival complexes : re-arrangement(5) and indexing. The same basic operations may be performed with the aid of electronic means. Both operations, whether carried out using the conventional or the automatic procedures, are based on the collection on forms of the data destined for the processing of conservation aids or of finding aids.

For the collection of data destined for automatic processing, a form is used that answers well-defined requirements, standardized in the base structure but differentiated according to the applications and, hence, reflecting the particular data processing level chosen for each archival group.

1.- Collecting form

The collecting form must primarly answer a number of requisites of a general character, which are in part the same as those of a conventional-type form. The peculiar requisite of automatic procedures is that the form must possess the flexibility needed for adapting itself to all homogeneous documentary sets in respect of both structure and typology. It is in fact a fundamental rule that all A.D.P. projects should proceed by groups of homogeneous fonds : the power of electronic means is such as to privilege to a large extent the search carried out on automatically processed sources and, therefore, programming must see to it that no inequalities in information occur in respect of the fonds belonging to the same search areas.

Furthermore, the form must be so arranged as to acquire all the homogeneous data required and sufficient for the chosen information level, with due regard to the documentary sets that are processed for a same data bank. Therefore, the lack of a typology of serial data in one of these sets in no way exempts from providing, in the form, the fields destined for it.

Eventually, the form must be capable of receiving, as in the case of the conventional-type forms, those non homogeneous data that are not regarded as relevant for the chosen data processing level.

In order to establish these general requisites in respect of each form and, hence, their final format, a procedure was established, at the end of which the form becomes operative and it will be thus possible to process the acquisition program on magnetic media of the data contained therein.

2.- Procedure for the formation of the collecting form

This procedure starts with a phase of analysis of the structure of the homogeneous archival groups destined for processing.

Homogeneity is appraised on the basis of the activity carried out by those who formed the fonds, of the chronological and territorial areas in which such an activity was carried out, of the institutional duties (namely, the assignements and fields of competence of the individuals or corporate bodies that formed the archival series), and of the manner in which such an activity was performed (which reflects the set-up of the body that carried it out).

The analysis of the structure takes place at the level of the archival entities or of single documents, depending on the chosen source of data collecting. In this analysis, such elements are considered that refer to the class of objects in respect of which the activity of the body that formed the archival series was carried out, to the typology of the persons that performed this activity and to the nature of the assignements on the basis of which they have acted, to the formal and juridical requisites that

concur in the perfection or in the completeness of the documents drawn up, and lastly, to the relationship existing between all the said elements.

Then the analysis is performed of the homogeneous data presented by each archival group, and of those of the non-homogeneous type, with a view to establishing the extent of data collecting and the extent of the allowed information kinds, as well as the retrieval aids that it is thought advisable to make available. A scarcity of the homogeneous data, when accompained by a wealth of non-homogeneous data, may lead to the decision of adopting a very pronounced extent of data collection and of taking the natural language of the texts; converserly, the presence of steady homogeneous data (of the serial type) may lead to adopt a pronounced extent of information kinds, and to adopt classes of descriptors as well.

On completion of this phase of analysis, a provisional collecting form, made suitable for the chosen data processing levels, is processed.

The second phase of the procedure consists in checking the provisional form on a suitable sample of textual sources. To make sure that this checking is valid, the extent of sampling must be made proportional with the extent of the data process to be performed. The choice of the samples, based on the previously made analysis, is carried out on the various record groups forming the archival group, and whithin same, by chronological bands, by territorial areas, by the fields of competence of the body that formed the archival group. If the source of data collecting is the single document, the items are sampled on the basis of the various elements presented.

During the work for the drawing-up of the provisional form, the pre-analysis thus made is actually checked. It is thus possible that new analysis are made, with ensuing modifications in the provisional form, and with further checking of same. In point of fact, the original form passes, as a rule, through a number of modifications during the checking phase. Moreover, the further analyses lead at times to an adjustment of the adopted levels of data processing. For example, classes of data may be excluded, which, on checking, proved to be non-homogeneous, thus narrowing the extent of data collecting; or one may act on the language, deciding to include classes of descriptors; it may be thought convenient to arrange more sophisticated retrieval aids.

On completion of this second stage of the procedure, the ADP level has basically emerged and the final form may be thus processed.

3.- Characteristics of the collecting forms

The collection forms are marked by a location code, whose function is to connect, in an univocal way, the data collected therein at their archival source. The code is conceived on the basis of the organization of the archival group to be processed, and must allow the accurate location of the information source.

The form is organized in such a way that each typology of the data to be collected is comprised in a

range of wider levels, divided according to the information procedures to be dealt with by the data collected therein, namely, re-arrangement, inventory, processing of retrieval aids. Within these general fields, the data may be grouped into sub-fields, according to the type of operation for which they will be used, or to the class of information to which they are assigned.

The organization of the form is not binding, however : in fact, the method chosen for the processing allows to utilize single data, irrespectively of the fields in which they were collected.

The collecting forms are of two basic types, oriented to the information procedures for which they are destined : re-arrangement and indexing.

The form that receives the data required for the re-arrangement and inventory of archival groups, does not exclude the utilization of these data for processing retrieval aids; it being an aid conceived for an orderly conservation of a documentary complex, it tends in fact to give birth to such an organization of the said set as to allow the retrieval (for operational or search purposes) of the archival items forming it(6).

On the other hand, since re-arrangement must be made at the level of archival entity, the source of data collection cannot be but the latter; hence, the language will prevalently be artificial and the classes of searches will prove to be quite limited. But a level of data processing which naturally is not much in depth, may be extended, thanks to the power and to the flexibility of electronic means, by acting on the retrieval aids. Therefore, a field was planned for this form for the collection of leads of arrangement, to be used for the formation of aids exclusively meant for search. These leads of arrangement are extracted from the area of the form relating to data which express the essential content of the archival entity namely, from entitling. Depending upon the criteria used for the drawing-up of entitling, it is possible that key-words of the source are found therein.

The form to be used for the making-up of retrieval aids may be applied only when the archival groups are already arranged. Although this form is oriented towards the collection of data in natural language, the presence is however envisaged of descriptors, such as for instance, the tabelled symbols, which, in the finding aids, occurr in standardized words belonging to previously established classes. In ti, the detection is always foreseen of the structural notations of the arrangement, which allow the appraisal of the retrieved information within the historical structure of the source from which they originate.

This type of form can be adapted to more or less in depth search levels, depending upon the source of data collection, upon the extent of data collecting, upon the ways of access allowed an upon the intermediate finding aids to be processed.

4.- Format of collecting forms

- form for arrangement and inventory
  Since the form is applied to the re-arranging procedure, its location code is coincident with the place provisionally given to the archival item. On completion of the arrangement operation, the

code will be made final. It will be connected to the provisional code or will replace it, depending on the convenience of carrying out the physical re-arrangement of the archival entities forming the documentary set.

The form is divided into the following fields :

A.- field meant for receiving structural elements of the archival group, namely, all those that, in the course of the pre-analysis and of the subsequent checking stages, emerged as possible leads of the primary arrangement, and that must be thus made the object of automatic analysis for the reconstruction of this arrangement. They may be in turn collected in separate subfields. As a rule, such subdivision is between chronological and topical elements; notations of the primary arrangement emerging with an appreciable frequency (extreme of placement, classifications); items relating to the institutional duties of the body that formed the archival group; words relating to classes of material objects of the activity carried out by the body itself (the classes normally envisaged are : toponyms, names of material objects).

B.- field meant for receiving notations for the processing of inventory. These are, of course, further notations in respect of those structural. They are usually collected into two separate subfields, one meant for entitling of the archival entity, the other for the automatic indexing of the annexes to the entity, by means of a tabelled symbol that indicates its typology.

C.- field meant for receiving data for the processing of aids for the retrieval of information items. Also in this case at least two subfields are envisaged, one for the collection of antroponym references and the other for exactly named references of other nature.

D.- unformatted fields meant for the collection of information on the data collected in the formatted fields of the form.

E.- unformatted fields meant for notes of an editorial type and for information items not specifically referred to a single record collected on the form.

- form for the making up of retrieval aids

The application of this form presupposes that the documentary set to be processed occurs arranged already. Thus, the location code coincides with the exact placement of the information item.

The form is divided into the following fields :

A.- field meant for receiving structural data of the archival group dealt with. These data usually are of a chronological, territorial and institutional order.

B.- field meant for receiving homogeneous search data, collected from the text and, hence, in the natural language. Each class of data occupies, within the field, a well defined subfield.

The aim is to collect leads of arrangement oriented towards the formation of intermediate finding aids and towards the processing, in interactive access, of retrieval aids of the quantitative type.

C.- unformatted fields meant for the notes to the data collected in the formatted fields of the forms.

D.- unformatted fields for notes of an editorial type and for information items not specifically referred to a single information unit collected on the form.

5.- Collection rules and criteria

Collection rules are oriented towards objectives that are common to any information science : never to cause information inequalities within the same search areas; to facilitate search through aids capable of avoiding noise or silence; to give the user the largest possible freedom of search as allowed by the adopted processing level.

A.- As concerns the standardization of information levels, the intervention is chiefly aimed at the unformatted fields, insofar as the very formulation of a single data-collection form for all documentary sets having a similar typology and structure, and the destination of the formatted fields for the collection of homogeneous data represent, upstream, a sufficient warranty of standardization.

The collection rules in respect of unformatted fields are formulated bearing in mind the analysis and checking made during the form making up procedure, but they are for the most part defined during the same data-collection operations, by means of automatic checking and controls on the decisions adopted, performing any possible revisions on the already acquired data. The latter operations may take place automatically, thanks to the almost simultaneous character of the operations of data acquisition on magnetic media.

Entitling is the operation that is most liable to subjective appraisals on the part of the system analyst, for the very fact that its formulation calls for a true professional skill. It is an established fact that the original entitling is collected, as a rule, whenever it presents all the necessary and sufficient elements for the understanding of the essential content of the archival item. Should this condition fail to materialize, the entitling processing is in the artificial language, but it should include the present original elements, provided that they are valid. Finally, if entitling is absent, it is entirely processed by the system analyst, in accordance with the archival rules. Since the leads of arrangement to be collected in the fields meant for processing of retrieval aids must be extracted from entitling, the maintenance of standardized information levels is ensured by the rules governing the latter's data-collecting rules. The data to be acquired for the entitling of the archival entities are established for each archival group during the procedure for the completion of the final form.

The typology of the notes to the data collected on the form, even though indicatively established during the preliminary analysis and checking, is defined during the data-collection; in point of fact, a checking of the provisional form - however comprehensive it may be - is hardly likely to highlight all the typologies of unformatted information that the documents might offer. According to the characteristics of the sources of data-collection, general rules may emerge from the initial checking on the language to be used in the notes - a natural language, as a rule, if the source of data collection is the text of the single documents.

The intervention in the formatted fields is mostly directed at the standardization of the open tabelled symbols. Other interventions are aimed at establishing well-defined criteria for the appraisal of the data to be ascribed to the different fields. It is a general rule, for the formatted fields meant for receiving the structural data, that these data should be collected in the natural language; no subjective intervention aimed at grouping together the data emerging from sources under descriptors is allowed, because the analysis for which such data are meant would prove off the track, since it would merely be an analysis of a possible logical arrangement that the system analyst has tried to superimpose.

B.- The rules aimed at facilitating search, preventing noise or silence, concern the language standardization. If the language is artificial, they consist in the standardization of the descriptors for eliminating synonims and homographs. Since, as previously pointed out, acquisition occurs almost simultaneously with collection, the procedure for the formation of the alphabetical list of descriptors may take place concomitantly with the data-collection., through the automatic processing of lists from time to time updated by the descriptors used in the various fields of the form. Each descriptor is accompained with the reference to the collection forms in which it was used. These lists are submitted to the working group that performs the data-collection, in such a way that the standardization of descriptors may take place in the light of the semantic value with which the words were used.

The intervention on the natural language is performed through the lemmatization of the key-words collected, in order to eliminate all grammatical homographs and to make the use of the finding aids easier.

The lemmatization is made at the time the key-word is collected in the formatted field.

As to the unformatted fields, and as for a full-text data-collecting one may decide to proceed to an automatic lemmatization(7) or to let the search be made through a mask during the interactive access. Another intervention envisaged on the language is the connection of the graphic variants; this operation is however performed during the making up of the aids for the utilization of the data bank.

C.- The objective of leaving to the researcher the largest freedom of utilization allowed by the adopted data processing level, materializes in the establishment of collecting criteria having the highest possible objectivity. This means : to limit the recourse to tabelled symbols that, as in the case of all descriptors, cause a limitation in search(8); to collect the key-words in the graphic variants in which they occur, chieffly to allow lexical searches; to adopt, to the maximum possible extent, a natural language in the unformatted fields meant for the notes; not to indicate in the formatted fields the missing data, even when they can be deduced with certainty, using for these the field meant for the editorial notes; to always indicate in the editorial notes the interventions made during the data collecting (dissolution of abbreviations, correction of material errors, etc).

All the above refers to the basic set of rules; a case list of the particular criteria of data-collection established in regard of each treatment, would be difficult to make; these criteria being often decided upon during the data collection itself.

CONCLUSIONS

As hinted above, the task of those who carry out an information science is that of preparing aids suitable for facilitating the retrieval of information items in the documentary sets without taking upon themselves the search tasks that are in duty of the user.

It is thus not at all correct to prepare specific finding aids, but one must always bear in mind all possible investigations that can be carried out, both when the choice of the data processing level is made and when rules and criteria of data-collection are established.

In this approach, the tendency exists to exploiting the advantages of the data-collecting form, by connecting it to a full-text data-collection. These advantages lie, in the first place, in the informational value assigned to each field meant for receiving homogeneous data, that leads to an appraisal of the specific information of which the datum collected is the carrier, and brings forth a distinction of homographs and a connection of synonims. Finally, the adoption of the form allows the lemmatization of words upon data-collection.

The aids that can be processed on the basis of the data-collection on forms, even when all possible searches are taken into account in the data-collecting rules and criteria, cannot however afford those in depth levels that can be achieved through the interactive access to the full-text of the documents where the word retrieved is displayed in its full context; but when this is accompanied by the data processing by means of forms, the structure within which data were collected as well as the aids that can be accordingly processed, represent a valuable grid, within which to direct a number of investigations(9).

NOTES

(1) By "natural language" we mean the language of the data collecting sources; this term is used merely in contraposition to that of "artificial language" (as used by the system analyst). Whenever the source of data collecting is the text of the original document, the natural language coincides with the textual language.

(2) It is the one that is expressed in the "entitling" (either original or formulated by the system analyst) of each archival unit.

(3) So are called those data for which the possibility is envisaged of being included among the finding aids as informational notations capable of receiving different kinds of arrangements (chronological, alphabetical, structural, logic, etc.).

(4) As concerns the normative thesauri applicable in archival fonds processing, please refer to : E. ORMANNI; *L'elaborazione automatica dei documenti di archivio*, Proceedings of the "First International Conference on Automatic Processing of Art History Data and Documents", Pisa, Scuola Normale Superiore, 1978, p. 147 and foll.

(5) The term "re-arrangement", as used here, means the taking-back of the arrangement to the original one, according to which the archival group was formed.

(6) Such conventional conservation aids as inventories are actually arranged and used in Archives, also for search purposes.

(7) As for the automatic procedure for the lemmatization of the words in Latin, please refer to : MINISTERO PER I BENI CULTURALI E AMBIENTALI - UFFICIO CENTRALE PER I BENI ARCHIVISTICI - GRUPPO DI STUDIO PER L'INFORMATICA; *Dimostrazione di applicazione dei mezzi elettronici alla ricerca d'archivio*; Rome, 28-30 September 1977; E. ORMANNI; *La ricerca automatica di documenti in un Archivio di Stato*; in "Data Report", 1/78, Milano.

(8) For example, whilst tabelled symbols are used for titles of nobility or clerical titles, for professional or office qualifications, codes are never used for arts and craft, in view of the peculiarity of same. In the data processing of field books, no codes were used to indicate the various types of crop, resorting, instead, to a data-collecting in the natural language and referring to the retrieval aids the making up of classes of crops.

(9) This data processing was adopted, even though according to different criteria, for the collection of the *Deliberazioni del Maggior Consiglio Di Venezia* (Libro d'Oro), and for the telegrams of the *Ufficio Cifra* of the *Direzione Generale di Polizia del Ministero dell'Interno,* kept at the *Archivio Centrale dello Stato.*