XMM-LSS
INTERNAL
REPORT n°8-Mi

May 10, 2011

Considerations about the XXL database and catalogue

The Milan proposal, and lessons learned from XMM-LSS

L.Chiappetti[1]

INAF, IASF Milano, via Bassini 15, I-20133 Milano, Italy

Abstract. This report provides supplementary information in conjunction with the presentation I prepared for the Bonifacio XXL meeting. It can be distributed to persons interested, particularly if they are not coming from the XMM-LSS collaboration.

Key words: XXL

1. Introduction

This document gives supplementary details to the material included in the presentation "*A proposal for XXL database and catalogue based on the XMM-LSS experience*" I prepared for the Bonifacio kick-off meeting of the XXL collaboration (May 2011), which will be given on my behalf by Marguerite Pierre. It is intended mainly for people which enter the XXL collaboration but were not part of the XMM-LSS one. XMM-LSS members will know already very well the details listed in section 2.1, but might wish to enter the discussion about "lessons learned" and suggests changes and improvements.

2. The XMM-LSS catalogue and database

2.1. Historical background

The allocation of responsibilities within the XMM-LSS collaboration assigned to INAF IASF Milano (formerly IFCTR CNR) the administration of the *master catalogue web site* and of the *database*.

In the next section we will give some practical information about the various components. Here we recall the addresses of the web sites, and supply some basic references.

- the original support web site is at `http://cosmos.iasf-milano.inaf.it/~lssadmin/Website/LSS/`
- the database web site is at `http://cosmosdb.iasf-milano.inaf.it/XMM-LSS/`

Most parts of those sites are publicly accessible. In particular the database web site has a *public* workspace which is accessible to everybody (with previous registration) and allows access to *published* results.

Published catalogues are currently represented by the (very partial) XMDS/VVDS 4σ catalogue (Chiappetti et al., 2005), which used an adaption of the Milan pipeline for point sources (Baldi et al., 2002), and by the XMM-LSS catalogue version 1, aka `XLSS`, (Pierre et al., 2007), which used the state-of-the-art version of Xamin (Pacaud et al., 2006) and covers about half of the XMM-LSS area with 3385 sources. The full area catalogue, reprocessed according to the latest version of Xamin (`2XLSS`) is in preparation (Chiappetti et al. , 2011), and the various steps towards it are documented by internal reports. It includes 6723 sources at the maximum depth (less if exposures are "uniformized" to 10ks, work in progress).

2.2. Breakdown

The XMM-LSS material mantained online in Milan can be broken down into components in different ways.
E.g. from the point of view of the data:

- raw Xamin catalogues (classified as *data products*)
- X-ray tables in the database
- other wavelength tables in the database
- multi-λ catalogues
- images, maps, spectra etc. (the typical *data products*)

A different breakdown may look at software components or at the way data is stored and accessed by them:

- the database user interface
- the database
- material outside the database

Other views are possible. We try to give a comprehensive though brief summary in the following sections.

2.3. Basic procedure

The current choice inside the XMM-LSS project for the processing pipeline for XMM data is Xamin version

Python 3.2. I presume details will be supplied at Bonifacio by the presentation of F.Pacaud, or could be asked to him or to N.Clerc. The reference paper for (an older version of) Xamin is Pacaud et al. (2006). For XMM-LSS the Xamin pipeline was run at Saclay, and its results were supplied to me in forms of *FITS catalogues* (or "cats").

The pipeline is run separately *per pointing* and *per energy band*. Therefore I received two files per pointing (soft and hard band). They include all tentative detections, inclusive of those with a poor detection likelihood, and for each of them give results computed according to *two possible solutions*, *as if* the source were pointlike and *as if* it were extended. An outline of the parameters included in the raw FITS cats is given in Pierre et al. (2007). Some other parameters are supplied separately (off-axis angles), or derived a posteriori (fluxes and uncertainties).

The first coarse step for what concerns the database is *ingestion* of the FITS cats, which *logically* supposes a number of sub-steps:

– one step is *band merging*: the independent detections in the soft and hard band have to be combined (but of course there are plenty of cases detected only in a single band).
– this step is combined with *classification* as pointlike or extended. If the detection meets the rules in a recipe (so called C1 and C2 cluster classification), which has been validated for the soft band but can be nominally applied also to the hard band, the source is classified as extended. For band-merged cases the extended/pointlike classification of the soft band prevails.
While extended sources are by definiion non-spurious, pointlike detections are flagged as spurious if the detection likelihood is below a given threshold. Therefore one can have: (a) sources detected as non-spurious in both bands; (b) sources detected as non-spurious in a single band and undetected in the other; (c) sources detected as non-spurious in one band with a spurious detection in the other; (d) sources detected as spurious in a single band.
– additional data like off-axis angles are inserted separately.
– finally position errors and fluxes are derived from other parameters.
– another step is *astrometric correction* which uses the SAS task `eposcorr` to apply small offsets to X-ray positions according to the position of optical candidates.

Pierre et al. (2007) remains the reference for the ingestion procedure, although the final catalogue in Chiappetti et al. (2011) will use a larger ($10''$) radius for band merging (as well as overlap removal, see below).

During the ingestion procedure, there is information present in the FITS cats, which is either considered irrelevant for the database, or is de facto discarded (e.g. the results of the extended source fit for a pointlike source or v.v.). To allow interested people to access it, the original ("raw") FITS cats are made available as *data products*.

By *data products* I mean data which is not *part of* the database but can be accessed (retrieved) *through* the database. Other X-ray related data products are images, exposure maps, contours, wavelet images, produced and used by Xamin in Saclay and hosted in a repository in Milan.
Note that no other intermediate results (e.g. event files) are stored as data products. Note also that the original ODF (once public) can always be retrieved from the ESA archives.

The next logical step for an X-ray catalogue is *overlap removal*, i.e. when the same celestial source is detected in overlapping pointings, only one entry shall be selected. In practice this step is *deferred* to a later stage: see the tables vs catalogues in section 2.5.1.

Usage of data in other wavebands is discussed in 2.6.

2.4. The database user interface

The original (dedicated) user interface for the XMM-LSS database was via a monolithic Java servlet, developed by Luigi Paioro as part of his master thesis. The current interface (i.e. the one which anybody can try at `http://cosmosdb.iasf-milano.inaf.it/XMM-LSS/`) is based on ~~**DART**~~▷ (see Paioro et al., 2008), which is a Tomcat/JSP solution used to support several other projects (e.g. z-Cosmos, VIPERS etc.) at IASF Milano.

We do not present here any screenshot to illustrate the look and feel (some will be shown in the Bonifacio presentation) but encourage everybody to test the real system at the URL given above registering in the public workspace.

We just list here the major features of the ~~**DART**~~▷ user interface:

– it supports personal *login* with user-driven registration. The user can also modify details of the account (like the e-mail address which is inserted in an user mailing list) and remove one's own account.
The user registers oneself onto the public workspace. Membership to other workspaces is activated on request by the database administrator after consent of the project PI.
– the user can select one or more *database tables* from the list associated to each workspace. See 2.5 for further details.
– the user can issue *simple queries* on the chosen table(s), which range from a complete retrieval of all data (not the best way to work, see 4.4), to a selection of some columns according to some simple criteria (in a simple sky region or according to simple criteria on columns).
– the user can also issue *advanced queries* exploiting a subset of `mysql` syntax. This includes sorting and lim-

iting the query, as well as usage of expressions among columns to create on the fly new columns or to select according to complex criteria.

- one can inspect the *results* visually or graphically or save them as text or FITS or VOtable files and later retrieve them.
- one can access the *data products* associated to the sources in the results, either one file at a time, or constructing on the fly compressed tar files including families (or the totality) of files.
- one can also store *often used queries* for later re-use.

There are other prototypal tools (not integrated in ~~DART~~▷) for a more flexible graphical representation, or for the visual overlapping of X-ray positions and counterparts in other wavebands onto thumbnail images in the optical or other bands. Some of these tools have been used in production by me or selected collaborators to support the identification work for XMM-LSS.

2.5. The database

Technically speaking, the bulk of the tabular data (i.e. excluding the data products, see 2.7, which can be physically bulkier in terms of megabytes but are stored as external files) is stored under `mysql`.
More precisely there are two databases (collections of tables) in `mysql` parliance. One is the *administrative data base* which is used internally by ~~DART~~▷ only and stores things like the names of tables and columns, their captions, units, priorities for ordering etc. etc. No further details will be provided here.
The other one is the *science data base* with the real data.

The user interface (~~DART~~▷) described in 2.4 above accesses both databases using the JDBC protocol, and under a single `mysql` user (which isolates any security concern). For maintenance tasks I use the line-mode `mysql` client manually or inside shell scripts (for typical table population I have more or less standard, repeatable scripts which as by-product produce also a log in HTML format). I can also use the line-mode client of `mysql` as a tool to issue complex queries useful for specific investigations. So far we did not want to give access via `mysql` client to remote users for security considerations.
A very limited number of specialized tasks is handled by C or C++ executables interfaced to `mysql`.

2.5.1. Tables vs catalogues

The science database contains everything which is visible to the users as well as everything which is hidden but necessary to present things to the users.

The typical entities in the database are *physical tables.* I use this name for simple tables, which contain list of sources in a given waveband or coming from a single point of origin, and which are usually *exposed* (i.e. not hidden) to users. I reserve the name of *catalogues* to something else (see further below), despite the fact that a physical table may correspond to the result of ingestion of several Xamin FITS cats (see 2.3), or to a subset extracted from a catalogue in another waveband (see 2.6), or e.g. to a list of NED or SIMBAD objects (see 2.7).

Not all columns in a table are necessary exposed. Some are hidden (not advertised) but can always be accessed if one knows their names.

A choice I made and strongly support is the fact that each record (row) in my tables has two special columns. One (seq) is a sequential numeric identifier, automatically generated, which univocally identifies the record. The other one (time) is a timestamp which records the moment the record has been modified.

Not all tables in the science database are exposed, and not all of them are plain physical tables.

- *correlation tables* are not exposed. They are simple two-column tables which store the association between the seq's in two tables according to a predefined criterion (usually proximity within a given radius, but this is not necessarily the only condition). When selecting *two* tables, the user can *choose* which of the existing correlation tables to apply. Their usage is strongly recommended as they definitely improves the efficiency (speed) of the queries.
- *views* are a `mysql` feature which allows to generate on the fly specific queries including a select condition. In their simplest usage they can be used to define subsets of existing tables which appear as normal tables.
- *GCTs* ("glorified" - or generalized - correlation tables) are not exposed, and extend the concept of correlation tables to n tables. They have at least n columns which are the seq's of objects (in the n tables) associated according to some criterion. They may contain other ancillary columns. GCTs are never accessed directly by the user (although their columns might be accessed if their name is known), but via the virtual table they support.
- *virtual tables* combine the usage of GCTs and views. They allow to select a choice of columns from n different tables (or expressions derived from their combination) for the objects associated by the underlying GCT. The n tables are called *member tables.* The columns defined in the view are called *virtual columns.*

A *catalogue* in our terminology is a virtual table, in which the first member table is privileged (in particular this is a band-merged X-ray table), and which includes only objects which are non-spurious and free from overlaps among different pointings (these conditions are achieved creating the appropriate GCT).

There are two basic types of catalogues:

- an X-ray catalogue has only 3 member tables, the band-merged X-ray table (for which only non-spurious overlap-free sources are chosen) and the two individual (soft and hard) band tables from which some columns are derived.

- a multi-λ catalogue has in addition other member tables (see 2.6). A particular X-ray source may have one, (none) or more *counterpart sets* which could be optionally ranked according to some criterion.

When convenient one can create from a catalogue other subset catalogues (*views within views*) according to particular criteria (although sometimes with a little performance penalty), e.g. if one wants to limit the counterparts of an X-ray source to just one (the best ranked).

2.5.2. What's in a name ?

Particular attention has been given to *naming of columns*, typically trying to use the same name for the same (or similar) quantities in different tables. However catalogues and physical tables may use different naming conventions.

The user shall be prepared to be flexible and know that e.g. a columns named Xseq in catalogue XLSS is the same as column seq in the first member table nov06.

Other care has to be given to *source names* (in the IAU format *prefix* J*hhmmss.s*±*ddmmss*). Coordinates shall be truncated, not rounded. Prefixes shall be registered with the IAU. Names shall not be altered if coordinates are updated. The full set of guidelines from the IAU is available at http://cdsweb.u-strasbg.fr/Dic/iau-spec.htx.

To prevent mistakes I registered with the IAU a prefix XLSS (and will soon register 2XLSS), which shall be used *only for sources officially published inside a catalogue* as well as prefix XLSSU which shall be used *for sources in advance of catalogue publication.*
We also defined an UDF (User Defined Function inside mysql) catname which can be used to generate the correct coordinate name. The virtual columns generating catalogue names in the various catalogues (Xcatname and alike) make transparent use of such function.

2.6. Non X-ray data

It is not obvious to estimate the respective amount of X-ray data vs data in other wavebands currently present in the Milan database, because I keep several obsolete or obsolescent releases, or data which was used in the past (for instance for the XMDS X-ray catalogue I used VVDS photometry, while now for XMM-LSS we use CFHTLS photometry, although we may use VVDS spectroscopy), but is likely that *in the database proper* the non-X-ray material exceeds the X-ray one by a factor 2-3.

For what concerns data products (see 2.7) the ratio is about 1:1, with 1519 Mb of X-ray data (mainly images and exposure maps, just 20 Mb of FITS cats) and 1426 Mb of optical and SWIRE thumbnails (the latter in up to 7 bands).

Procedure-wise the activity about non-X-ray data at large can be classified as:

- data ingestion
- correlation with X-ray tables
- optical identification
- optional retrieval of data products

Data ingestion depends on the data source. For small published catalogues material might have been supplied in advance by the author (if it was a within the framework of the XMM-LSS collaboration) or I extracted it from the paper LATEX source (there are database tables for our own papers and some other ones). In all these cases the entire catalogue (which is public) is ingested as a database table. Similar considerations hold for miscellaneous tables (like the poorly used "spectroscopic followup" one) or for low density catalogues (like SIMBAD, NED or USNO). For the latter, one can populate a table via queries returning all objects in a given X-ray pointing and later do a correlation with our X-ray source position on a narrower radius.

For larger catalogues one has to make *data right* and *size* considerations. Data rights are usually regulated by a Memorandum of Understanding and restrict the right of access of the XMM-LSS consortium to some authorized subset of the data. At the same time size and disk space considerations suggest not to keep all the data, even if publicly available, online in our database.

So the ingestion usually requires a *preliminary correlation* with some coarse radius (of the order of 9-10$''$ about non-astrometrically corrected X-ray position).

When possible (this was the case e.g. for UKIDSS, GALEX or the Spring 05 release of SWIRE) the extraction of the subset was done at the respective public archive site (WSA, MAST or IPAC) uploading a list of X-ray source positions, running *there* the correlation, retrieving a (usually FITS) file and ingesting it in our database.

In other cases when the data was not public we obtained from collaborators, members of both consortia and entitled to do it, a list of optical or IR sources was obtained, ingested it in a temporary table, and deleted it after we extracted our own coarse subset within 9-10$''$ using our own correlation program. This way was followed e.g. for various releases of CFHTLS (optical photometry) or SWIRE (including the latest DR6).

A problem related to these catalogues concerns overlap removal of duplicated sources observed in adjacent, partially overlapping pointings, and eventually other cleanup tasks. These will be discussed in 4.2.

After ingestion we generate *correlation tables* between the X-ray table of interest (i.e. the latest release) and the ingested table using a smaller radius (typically 6$''$), using our own very fast correlation program.

One can get a flavour of the tables present in an of the workspaces of the Milan database by consulting the list (and links) on `http://cosmos.iasf-milano.inaf.it/~lssadmin/Website/LSS/List/`.

2.6.1. Identification

The generation of a multi-λ catalogue requires an at least tentative identification of the X-ray source with sources in other wavebands. The full description of the current procedure is outside of the scope of this report (it is described in other XMM-LSS internal reports and will be summarized in Chiappetti et al. 2011).

For the latest catalogue we used in the optical band the CFHTLS D1 and W1 fields (release T004), complemented by three pointings (the "ABC fields") obtained by M.Pierre under GO time in the northernmost part of the XMM-LSS area. In the IR band we used the SWIRE DR6 data. In the NIR band we used the UKIDSS DR5plus release (with a rather partial coverage). In the UV we used the GALEX GR4/5 release.

In addition there are correlations (not integrated within the catalogue) between the X-ray position and NED, SIMBAD and the USNO A2 bright object catalogue.

The *pre-identification* consists in the creation of a GCT (see 2.5.1) which correlates X-ray sources with CFHT, SWIRE, UKIDSS and GALEX. The procedure is run *incrementally*, i.e. first I create an entry for each (X-ray, CFHTLS D1) couple within $6''$. Then I correlate X-ray positions with CFHTLS W1/ABC within the same radius, as well as within $0.5''$with the eventual D1 object. This might result in the generation of an edited triplet (X-ray, D1, W1), or the addition of a new triplet (X-ray, null, W1). Then I do the same thing for SWIRE, then UKIDSS, then GALEX (with radii of $1''$, $1''$ and $1.5''$).

This generates a list of n-uples, or *tentative counterpart sets*. At this stage for each band (with a counterpart) I compute a chance probability (also stored in the GCT and later available via the database) based on the X-ray-to-band distance r and the density $n(brighter\ than\ m)$ i.e.

$$probability = 1 - exp(-\pi\ n(brighter\ than\ m)\ r^2)$$

Such probability can be used for a preliminary ranking, considering *good* a probability < 0.01, *fair* one between 0.01 and 0.03, and *bad* one > 0.03.

To finalize the *ranking* a number of other empirical but objective criteria were used along with the combination of the probability-based preliminary ranks in the 4 bands (optical, IR,NIR and UV), like the fact the counterpart is the only one, is the brightest and closest etc. or biases in favour of SWIRE or optical counterparts, or the fact the ratio of the best probabilities between different counterpart sets is greater than 10, etc.

This may result in some (most) of the counterpart sets being rejected altogether, or of an X-ray source having a single preferred counterpart, or of an X-ray source having more possible, *ambiguous*, counterparts (in some cases with one definitely preferred, in other cases just nominally better).

It is important to stress that visual inspection (using one of the additional tools) of the counterpart sets over an optical (or IR) thumbnail image has been very useful to assess not only ambiguous cases, but also other suspicious cases like crowded fields, or cases of very bright sources which may be saturated or even not present in some of the photometric catalogues. In some cases this requested manual adjustments to the initial identifications.

2.7. Data products

We call *data product* any kind of data which is stored in a file or an external resource (i.e. *not* in a database table) but is accessible *from* the database as a result of a query.

We do not provide a comprehensive list, but just list categories of data products.

- X-ray associated data products include the FITS cats, X-ray images (normal and wavelet), exposure maps, ds9 contours.
- data products associated to optical or IR catalogues are thumbnail images around a given source, either in FITS or PNG format.
- optical spectra, or SEDs (as resulting from photometric redshift work or alike) can also in principle be supplied as data products.
- a particular kind of data products associated to sources in an X-ray catalogue are textual comments.
- one can handle as data products also resources like the web pages of NED or SIMBAD associated to counterparts of our sources.

- some data products might be associated to a pointing. Typically X-ray data products (except textual comments) are associated to the X-ray pointing.
- other data products are associated to the individual object. E.g. a radio map may be associated to an entry in a radio catalogue, but optical and IR thumbnails, as well as textual comments, are associated to an *X-ray* source (sic!).

- the actual association occurs via a database column. The locator of the data product has a fixed template with a variable part (which might be a filename, part of it, or a directory name) which is the value of the database column.
- the locator is a sort of URL which can be of `file` or `http` type.
- a `file` data product necessarily resides on a disk accessible from the database server.

- a `http` data product may reside also on an external server. In principle it could even be produced on the fly by a CGI script. Although I encouraged people to mantain data products on their servers without transferring them to Milan, de facto all data products are currently on local web servers except the pointers to the NED and SIMBAD web pages.
- a data product can be mandatory if it is always present for all possible values of the associated column.
- a data product may be optionally present, conditionally to the value of a flag contained in another database column.
- there is no need of a flag column for an optional data product, if it is of the `file` type. In this case DART▷ can automatically probe the file existence.
- however if the data product is of `http` type, the flag columns is necessary to prevent runtime error when trying to access the product.

Some data products (typically all the X-ray ones and the PNG thumbnails) are supplied by Saclay. Other data products may be supplied by other collaborators. Optical and IR FITS thumbnails were retrieved from public archives elsewhere (CADC for CFHTLS and IPAC for SWIRE) by me, using the provided cutout facilities, and stored in Milan after renaming according to the needed naming convention. I did not retrieve any UKIDSS FITS thumbnail because they use a WCS in an unusual (ZPN) projection not supported by my additional tools. Of course NED and SIMBAD pages reside on the respective site (and are accessed building an URL from a locally stored identifier).

3. The proposal(s) for XXL

In this section I present the proposal for participation to XXL activities of IASF Milano for what concerns the database and catalogues. From the point of view of manpower such proposal concerns essentially only myself.

One should remember that one of the two sky areas of the XXL (the one near RA=2 hrs) will be surrounding the area covered by the XMM-LSS, and that the data in the `2XLSS` catalogue as they will stand after publication (or in a revised form, if the Saclay pipeline is updated) will form integral part of the XXL survey.

3.1. The least effort case

A minimum effort will be required if XXL data are managed in a way extremely similar to XMM-LSS ones.

In this case I could just create an *XXL workspace* under the same DART▷ environment currently used by the XMM-LSS. Members of the XXL collaboration will be given access to such workspace. The interface will be the same one used currently for XMM-LSS i.e. the one at `http://cosmosdb.iasf-milano.inaf.it/XMM-LSS/`. A corollary is that *all* database tables will be stored inside the same `mysql` database (by all I mean both data tables and administrative tables).

It is TBD how to handle other information, like the logs of the operations done or ancillary information on the progress of observations (for XMM-LSS currently mantained semi-manually on the other site).

3.2. The reference case

The solution above is somewhat "politically inelegant" as it "hides" XXL inside the XMM-LSS database site. A more elegant solution, requiring no more *maintenance* effort *after an initial setup* could be the following.

A new DART▷ installation is made for XXL. This might require a una-tantum intervention by my colleague Luigi Paioro (DART▷ author). I will then install a new logo and home page, eventual ancillary pages, and mandatorily a new `mysql` administrative database (separate from the XMM-LSS one). The modality of access will be the same as the one used for XMM-LSS, only logos, colours etc. will be different, and only XXL-specific workspaces will be accessible.

The actual `mysql` database (containing data tables) will be instead the *same* used by the XMM-LSS (a few data tables will be shared between the two, while most of the others will be used only by one *or* the other as instructed in the relevant administrative tables).

3.3. Do it elsewhere ?

If some other institute or team with adequate resources and manpower is wishing to mantain the XXL database on their site using DART▷ and `mysql`, it will be possible to arrange for an initial setup of a DART▷ installation on a suitable machine (this will require a substantial intervention of Luigi Paioro), and for the subsequent exchange of the relevant know-how (via meetings with me, and subsequent e-mail support).

What I tend to exclude, or at least consider subject to further negotiations, is the implementation of new features within DART▷. DART▷ is used at IASF Milano by other projects, and the needs of XXL (or even XMM-LSS) might not always be compatible. This statement applies also to case described in 3.2.

3.4. Do it otherwise ?

If some other institute or team with adequate resources and extensive and competent manpower is wishing to take charge of the XXL database on their site using software of their choice, it will be possible to arrange for the initial exchange of the relevant know-how and experience (via one or more meetings with me).

4. Lessons learned and suggestions

I collect here miscellaneous items based on the XMM-LSS experience, like possible critical points or improvements or changes to the procedures used so far, as well as open questions due to differences between XXL and XMM-LSS, or just to issues which might be unknown to me.

The various items are marked with a capital letter (from (A) onwards) with consecutive numbering throughout the various subsections, in order to allow easy reference in discussion

4.1. The pipeline and X-ray data

Most of these items (**A** to **H**) concerns more the Xamin pipeline or the ingestion phase than the database itself.

(**A**) the Xamin pipeline could be improved to give errors on count rates and fluxes which are presently missing

(**B**) the Xamin pipeline could be improved to make the choice whether a source is extended or pointlike, instead of deferring it to the phase of ingestion into the database

(**C**) it would be highly desirable if the Xamin pipeline runs simultaneously on both energy bands. This will eliminate the need of band merging in the database ingestion phase, and could possibly always give a count rate measurement or at least upper limit in both energy bands

(**D**) if the last two requests are met, the Xamin pipeline should also be able to select directly non-spurious sources (likelihood above threshold). Spurious sources will not enter the database at all.

(**E**) the Xamin pipeline or the preliminary SAS tasks should be improved in order to cope with adjacent pointings. Currently detection is run independently on each field, and there is no exploitation of the sum of exposures in overlapping pointings. On the contrary overlap removal forces to choose one of the various detections. For XMDS (pipeline according to Baldi et al. 2002) I experimented with stacking of local data (which allows to improve statistics and decrease errors, although not to go deeper). I also did some experiments running a similar pipeline on SAS-mosaiced data (however on repeated fields on the same pointing, not adjacent ones). The recent improvements to the SAS in support of mosaic mode should be assessed, particularly in conjunction with the strategy and timeline of XXL observations.

(**F**) the tradeoff between exposure uniformity and maximum depth shall be discussed. A satisfactory solution to the previous request will allow to maximize the flux depth of the survey. On the contrary for XMM-LSS a recent decision was to privilege exposure uniformity at the price of cutting longer exposures to a maximum length of 10ks.

(**G**) I need more information on the observing strategy of XXL repeated observations of the same field. This has some impact on the timing of ingestion of data in the database and release of the data. In the past new releases of table families or catalogues coincided with the (re)processing of entire AOs. If one desires a quicker release for XXL, following the release of ODFs according to the "legacy" policy, one has to consider the interference with the removal of overlapping sources in fields (re)observed in the future ! In XMM-LSS parliance, "physical tables" could be easily updated adding new pointings with no harm, but "catalogues" will have to be redone, and overlapping removal might remove or replace sources !!

(**H**) I stress the importance of the *stability of the interfaces.* In the past, due to the ingestion occurring every 1-2 years in coincidence with the AO "rounds", and also to the turnaround of postdocs in Saclay and to the upgrades to the pipeline, I received somewhat different families of FITS cats at each new release, and had therefore to adapt the ingestion scripts and/or database table layout.

(**I**) Due to the fact new data came in somewhat unpredictably (and also to changes in the versions of `mysql`), the ingestion scripts weren't planned in advance or stable, but adapted or developed case by case. I tried to use a common strategy (e.g. logging the output into HTML pages), but each time I had to cope with a new input catalogue or release I had to spend time in editing and testing new scripts, which, also in conjunction with the limited manpower, causes delays.

4.2. The optical catalogues

I presume that data in other wavebands, as in the past, could be derived from different sources: (a) public archives (all-sky or not); (b) large programs with which we have a memorandum of understanding; (c)

(**J**) what is the coverage for the XMM-LSS area for what concern optical, IR and NIR (I expect GALEX is available) ? Does CFHTLS data exist ? Are we going to use other CFHT observations ? or other optical data ? What about Spitzer ? and UKIDSS ? anything else ?

(**K**) what kind of optical, IR and NIR data will be available for the BCS area ? Will they be consistent with those available for the northern area (e.g. in terms of magnitudes or energy bands) ? Is this advocating for two separate multi-λ catalogues or can one have a common one ?

(**L**) other bands or catalogues to be considered ?

(**M**) what about spectroscopic followup ? In the past these data were rather inhomogeneous, or came in quite late for ingestion (or not at all).

(**N**) validation and uniformity of optical (and non-optical) data is important particularly for photometric redshift and SED computation. Uniformity in the choice of apertures, normalization of magnitudes and any other aspect should be preferrably managed by a dedicated team or WG. It has to be clarified the level of interaction with the database as working tool.

(O) what is the availability of optical or IR images (and cutout services to extract thumbnails) ? One shall also evaluate a tradeoff between PNG and FITS format. Also ex-officio as IAU FITS WG vice-chair, I strongly endorse the usage of FITS thumbnail images (with proper WCS) which allow flexible display and usage with common astronomical utilities.

(P) considerations on the stability of interfaces and their impact on the ingestion scripts are similar to the ones raised in items **H** and **I**.

4.3. The identification

The following items possibly suffered in the past of a lack of manpower, and could be addressed in the context similar to the one mentioned in item **N**. Interaction with the database (also with write access !) and development and usage of additional tools interfaced with the database should be considered.

(Q) The empirical procedure for preliminary identification presented in 4.3 shall be discussed and re-assessed, particularly in conjunction with differences in the input catalogues.

(R) Also the ranking procedure mentioned in the same section shall be discussed. One particular item concerns the density $n(brighter\ than\ m)$ for the different catalogues, which I derived myself in a quick and dirty way, but which ideally requires the competence of a (different) person from the team which prepared the original catalogues. A similar experience is required to make sure the input catalogues are free from artifacts and duplicated sources.

(S) I am convinced that a fully automated procedure will not be possible, and that irregularities which are unavoidable in any catalogue will require a stage of visual inspection and validation (of course the database can assist in locating the cases worth of a prioritary inspection). This item raises also a concern about manpower.

4.4. Database usage

The consideration below are particularly applicable in the case the proposals presented in 3.1 or 3.2 are approved (those sections contain also considerations about limited manpower and limited software changes).

(T) Based on the XMM-LSS experience so far, I believe that introducing too restrictive data rights (by which certain users can access only certain tables), and enforcing them by a proliferation of workspaces (which were "invented" exactly to cope with such requirement), is not positive. It limits the way users can fruitfully exploit the data, places a burden on the database administrator, and in general slows down the work. De facto some of those workspaces ended to be never used. I suggest all XXL data are placed in a single workspace.

(U) Although the database interface does not give access to the full capabilities of the `mysql` line mode client (which I do exploit), it allows to do efficiently a lot of things (in selecting flexibly subsets of the data, in repeating queries after a database update, etc.). However it might be possible that putting effort in improving the user interface might be a waste of time, since many users seem to prefer to just use once the "GET ALL" button (for that we do not need a DBMS ! an ftp server will suffice !) and bring data home and then work with their own tools.

(V) On the other hand I must admit that I am VO-skeptical. I am not convinced of real usefulness of Virtual Observatory tools and in general I prefer to use my own tools, or those which I know best. I note also that VO is intended for access to fully public data, and therefore it might not be suitable when an user authentication (even a very simple one like "I am a member of XXL") is required. ~~DART~~▷ was built to produce VO-compliant output, but I think those features were never used. I note also that, if further VO requirements will arise, I will not have adequate competence and will require consultancy.

References

Baldi, A., Molendi, S., Comastri, A., et al. 2002, Ap.J, 564, 190

Chiappetti, L., et al. 2005, A&A, 439, 413 (XMDS/VVDS 4σ)

Chiappetti, L., et al. 2011, in preparation (`2XLSS` paper)

Pacaud, F., et al. 2006, MNRAS, 372, 578 (Xamin paper)

Paioro, L., Chiappetti, L., Garilli, B., Franzetti, P., Fumana, M., & Scodeggio, M. 2008, Astronomical Data Analysis Software and Systems XVII, 394, 397

Pierre, M., et al. 2007, MNRAS, 382, 279 (`XLSS` paper)