VLT Archive-based Papers scooping PI/CoI

1. Executive Summary

The 4217 VLT programs used in refereed papers published between 2008 and 2017 have been considered. For each of them, the first paper was characterized as “PI/Cols” (90.0%) or archive paper (10.0%). Of these programs resulting first in an archive paper, 131 programs (3.1%) are used first in an archive paper, then in a paper published by the PI/Cols; in other words, the original proposers have possibly been scooped (see Fig.1).

However, the time elapsed between the data release date and these first publications indicates that only 2 papers (2% of the “possibly scooped programs”, 0.05% of all programs) have a publication date close or before the time when the data became public. In both cases, it is plausible that the data were not actually used in the paper but constitute a citation error. Of the 129 papers published after the data became public, it is interesting to note that 65 programs (50% of the “possibly scooped programs” or 1.5% of the total) resulted in archive papers published at least 2 years after the data became public (i.e. typically at least 3 years after the data were acquired), and 99 programs (76% of the “possibly scooped programs” or 2.4% of the total) 1 year after the data became public. The distribution is shown in Fig.2. Overall, the current data proprietary period policy is working as intended.

An alternate policy could release the frames when the program is completed, making them public one year after. While such policy would have significant impact (which are beyond the scope of this document), the present study shows that only 4 programs would have been scooped, i.e. 0.09% of the programs, over the past 10 years. The distribution is shown in Fig.3.

Overall, archive research scooping the work of the PI/Col was not an issue over the period from 2008 to 2017. As the fraction of archive paper is increasing, this study could be re-done in a few years to verify the validity of this statement. In the meantime, archive research scooping PI/Col science does not justify changing the data proprietary period policy.
VLT Archive Papers scooping PIs

Fig.1: Distribution of the authors of the first paper on VLT data, based on 4217 programs from 2008-2017

Fig.2: for the 131 programs in category 3 (“First Paper Archival, PI/CoI possibly scooped”), histogram of the time $dT$ [y] between the first frame released publicly (after proprietary period) and the publication of the archive paper. This corresponds to the current proprietary period policy. The 2 papers highlighted in red correspond to incorrect data citation. The median delay is 2.0 years, i.e. typically 3 years after the first frame was acquired.

Fig.3: histogram of time $dT$ [y] between the date of public release of the last frame of a program and the submission of the first resulting archive-based paper. Negative values do not necessarily correspond to “scoops” under the current proprietary time policy.
2. Introduction

The Users’ Committee inquired on the suitability of the data proprietary period to protect the PI/CoI of proposals against archive researcher scooping their results, and proposed shifting the start of the proprietary period (and therefore the data delivery to the proposers) to the completion date of the run for normal proposals. As this would have significant impact on the community—in particular by preventing the proposers to evaluate their data before the completion of the run, the possible impact of the proprietary period on publications was evaluated.

This study is based on tables provided by the ESO library (Uta Grothkopf and Dominic Bordelon) using the ESO Telescope Bibliography TelBib and input from USD (Marina Rejkuba) and others. The underlying data are in a spreadsheet attached to this document in PDM.


The first step was to get a list of VLT Program IDs used in TelBib papers published 2008-2017. Then for each of these Program IDs, we queried a list of papers along with their Archival use of the program (Y or N), sorted by publication date (defined on the month level). The paper(s) with the earliest publication date were designated as “first paper(s)” of that program, and then the Archive column was used to categorize the Program ID into one of the four categories listed below.

<table>
<thead>
<tr>
<th>Data usage in program’s first paper</th>
<th>Number of programs</th>
<th>Fraction of Programs</th>
</tr>
</thead>
<tbody>
<tr>
<td>First paper by PI/CoIs (Archive = N)</td>
<td>3796</td>
<td>90.0%</td>
</tr>
<tr>
<td>First paper Archival; PI/CoIs never published</td>
<td>277</td>
<td>6.6%</td>
</tr>
<tr>
<td>First paper Archival; PI/CoIs possibly scooped</td>
<td>138</td>
<td>3.3%</td>
</tr>
<tr>
<td>Simultaneous Archival/non-Archival first papers</td>
<td>6</td>
<td>0.1%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4217</strong></td>
<td></td>
</tr>
</tbody>
</table>

The first category (First paper by PI/CoI, 90%) is the expected, nominal outcome of a programme.

The 6.6% of programmes resulting in a publication by another group, but not by the original PI/Col are difficult to evaluate: It is possible that for a fraction of them, the archive paper constituted a negative incentive for the original group. However, a survey on non-publishing programmes (Patat et al., 2017, Messenger 170, 51) indicates that about half of all programmes do not result in a publication by the PI/Col. The survey lists a series of option why

---

1 http://telbib.eso.org/
the programme did not yet result in a publication (the most important one being “we are still working on it”). A subset of 12.2% of non-publishing programmes invoked “other reasons”, which include, among several other options, “Other, quicker result by another group” (i.e. scooping). Hence, only a minority of non-publishing papers are expected to have been affected by archive scooping: fewer than 12.2% of the 277 programs that did not result in PI/Col publication, i.e. <34 programs.

The 138 programmes (3.3%) resulting in a first publication by archive researchers then by the PI/Col group constitute the core of “potential archive scooping”, which are evaluated in more details below.

4. Analysis of the possibly scooped programs

The 138 programs where the first paper is archived-based have been considered in more details.

4.1 Preliminary triage

A first triage step estimated roughly the delay $dT$ between when the data became public and when the archive paper was published. This was done crudely using

- Data becoming public: one year after the beginning of the Period of the program. This assumes that data were acquired on the first night of the period and can be wrong by up to 6 months (for data acquired on the last night of the period) or more (for observations carried over to subsequent periods).
- Publication date: 1 Jan. of the year of publication. This can be wrong by up to 1 year (for papers published on 31 Dec.).

Overall, this crude estimate can therefore be wrong by up to 1.5 year. This error is asymmetrically underestimating the time delay. The resulting histogram of “scoop time” is presented below in Fig.4. This is the histogram that was presented to UC in 2019.

![Scoop Triage](image)

**Fig.4:** preliminary histogram of time $dT$ [y] between the approx. release date of a dataset and the resulting archive-based paper.
4.2 Scooping under current proprietary period policy

The release date for the first and last frame of each of the 138 programs were extracted from the Archive. These dates properly take the proprietary period which is shorter than 1yr for some programs (2 of the 60 are public surveys). The "received" and "accepted" date of the archive-based paper were recorded.

To evaluate possible "scoops" against the current proprietary period policy, the following scoop time is evaluated: $dT = \text{[paper received date]} - \text{[release of the 1st frame]}$.

Eight programs are flagged as potentially scooped under the current proprietary period policy; for the others, the archive-based paper was published after the data were made public. They are listed below.

<table>
<thead>
<tr>
<th>Program ID</th>
<th>Data Archive</th>
<th>Publication</th>
<th>Scoop tests [years]</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>68.A-0544</td>
<td>2003-Mar-05</td>
<td>2003MNRA:S.342:439</td>
<td>2002-Dec-03 2003-Feb-13</td>
<td>-0.3 -0.1 Probably citation error</td>
</tr>
<tr>
<td>078.A-0003</td>
<td>2008-Apr-05</td>
<td>2008A&amp;A...487:583B</td>
<td>2007-Nov-14 2008-Jun-11</td>
<td>-0.4 0.2 Citation error</td>
</tr>
<tr>
<td>079.A-0303</td>
<td>2008-Jul-19</td>
<td>2008A&amp;A...487:583B</td>
<td>2007-Nov-14 2008-Jun-11</td>
<td>-0.7 -0.1 Citation error</td>
</tr>
<tr>
<td>079.A-0656</td>
<td>2008-Sep-17</td>
<td>2008A&amp;A...487:583B</td>
<td>2007-Nov-14 2008-Jun-11</td>
<td>-0.8 -0.3 Citation error</td>
</tr>
<tr>
<td>080.A-0288</td>
<td>2007-Dec-11</td>
<td>2008A&amp;A...487:583B</td>
<td>2007-Nov-14 2008-Jun-11</td>
<td>-0.1 0.5 Citation error</td>
</tr>
<tr>
<td>081.A-0334</td>
<td>2009-Apr-06</td>
<td>2008A&amp;A...487:583B</td>
<td>2007-Nov-14 2008-Jun-11</td>
<td>-1.4 -0.8 Citation error</td>
</tr>
<tr>
<td>089.C-0494</td>
<td>2013-Jun-10</td>
<td>2013ApJ...776:4N</td>
<td>2012-Nov-15 2013-May-30</td>
<td>-0.6 0.0 Probably citation error</td>
</tr>
</tbody>
</table>

Paper 2008A&A...487:583B, potentially scooping 6 programs, deserves additional investigation. This paper states “The data basis of our analysis are 103 optical spectra of high-redshift QSOs and AGN obtained with the Ultraviolet and Visual Echelle Spectrograph (UVES) between 1999 and 2004 at the ESO Very Large Telescope. These spectra, observed for various purposes by various groups, are publicly available in the ESO data archive”. As the actual list of programs is not included in the paper, the authors were poked by ESO library with a request to provide the programs they used. It appears that this list was produced post-facto using the archive status at that time, and does not reflect exactly what was used. In addition to the 6 programs listed, this paper also cite a 7th run (080.A-0014) at the limit of plausibility.
runs have been removed from this study, resulting in 131 valid runs (which are those reported in the executive summary).

Paper 2003MNRAS.342..439S is a study of z~6 galaxies in the Chandra Deep Field South using HST ACS data. Low redshift interlopers were identified using ESO Imaging Survey data released as part of the GOODS programme. They don’t mention 60.A-0544 in the body of the paper, but list it in the archive acknowledgement. It is unclear whether they actually used that dataset.

Paper 2013ApJ...776....4N publishes a planet search around 70 young nearby stars with high-contrast imaging with Gemini NICI. One target discussed in this paper was observed by 089.C-0494 (Kurtev et al): HR6037, made public in Jun.2013. They indicated they “retrieved the archival 2004 and 2006 VLT data used by Huelamo et al 2010”... but do not mention the data by Kurtev et al, which were not available before they published. It is therefore likely that this is a citation error.

Overall, on the remaining 8 programs potentially scooped, 6 are definitely caused by citation error, one is a plausible citation error, and for the last one it is unclear whether the data were used at all.

The Figure 2 (in the executive summary) shows the histogram of the time dT between the public release of the paper and the archive paper submission date. The two doubtful cases discussed above are highlighted in red.

Overall, the only two papers with conflicting negative dT are likely citation error. All the other papers were submitted after the data were made public. Of these, only 13 were submitted in the first 6 months after the data were made public, the others were submitted later, with a median dT = 2y after the data were made public (that is, typically 3 years after the data were acquired).

4.3 Scooping under different proprietary time policy

The User’s Committee pointed out that under the standard current proprietary period policy, each individual data frame is made public one year after it is acquired (that period can be adjusted for certain programs). For programs extending over a significant period of time, this implies that the PI/CoIs will have a reduced proprietary time over their complete dataset. In some cases, the first frames can be made public even before the program is completed.

They requested that ESO considers tying the public release of the entire dataset to the date of the last frame. This could have additional consequences, for instance making the data available to the PI/Col only once the run is completed, thereby precluding them to perform a preliminary analysis of the partial dataset and potentially refining their strategy. The discussion of these consequences is beyond the scope of this study.

In order to investigate the potential effect of such an alternate proprietary time policy, the following scoop time is evaluated: dT = [paper received date] - [release of the last frame]. As this concept does not make sense for Large Programs, they have been removed from the sample. The histogram of dT over the remaining 121 programs is presented in Fig.3 (in executive summary).
### Program ID | Archive | Scooping Paper | Publication | Scoop tests [years] | Comments 
--- | --- | --- | --- | --- | --- 
| 073.D-0670 | 2004-May-14 | 2005-Aug-03 | 2005MNRAS.359..906H | 2004-Nov-29 | 2005-Feb-12 | 0.5 | -0.7 | PI/Col and authors overlap 
| 080.A-0411 | 2008-Sep-15 | 2009-Mar-04 | 2009ApJS..183..244N | 2009-Apr-09 | 2009-Jun-23 | 0.6 | 0.1 | PI/Col and authors overlap 
| 083.C-0837 | 2009-May-01 | 2010-Jul-27 | 2010A&A..521A..18G | 2010-Feb-10 | 2010-Jun-08 | 0.8 | -0.5 | PI/Col and authors overlap 

The authors of 2005MNRAS.359..906H overlap with the PI/COI of 073.D-0670 (Benetti, Turatto); one author of 2009ApJS..183..244N overlap with the PI/Cols of 080.A-0411 (Rosati); the authors of 2010A&A..521A..18G overlap with the PI/Cols of 083.C-0837 (Guarcello; Micela; Peres; Prisinzano). These papers should not have been reported as “archive paper”, but as PI/Col.

The remaining 4 papers would indeed have “scooped” the PI/Cols under the considered alternate proprietary time policy.

In conclusion, changing the proprietary time policy to protect all the data of a program until one year after the program is completed would have prevented 4 programs to be scooped by archive papers, that is 0.09% of the programs over the past 10 years.

-end of document-