Chapter 7

Summary and Outlook

Billions of astronomical objects are detected with modern telescopes, for which hundreds of parameters are quantified. New techniques are necessary for interactive exploration of these large datasets. *Query driven visualization* is introduced in this thesis to achieve the required scalability while retaining flexibility and facilitating collaboration.

Research on the evolution of galaxies and their interaction with the environment highlights the requirement of using such large datasets. A large number of galaxies are necessary to discover relations between their properties and to subsequently find exactly those galaxies that do not follow the detected trends. Furthermore, flexibility in quantifying the properties of galaxies is required, because many of them can be estimated with different methods.

Visualization of these large datasets boils down to the question of how to show terabytes of survey data on computer screens of only a few megapixels. We tackle this scalability problem by turning it around and phrasing the question from the perspective of the desired result: Given a one megapixel screen, what needs to be done to visualize this one terabyte of data?

Starting from the desired end product, it is possible to automate the tasks required to achieve the necessary scalability, because at every step it is known what the end result has to be. In this thesis we research how astronomical catalogs should be handled to accomplish this scalability and we design mechanisms to automatically discover, create and process catalogs in a scalable and flexible way.

Query driven visualization is a technique that allows scientists to discover existing datasets and create new datasets by requesting data directly from within the visualization by using these mechanisms. This allows scientists to interact with their data in a conceptual way and allows them to focus on *what* they want to do with the data, because *how* this is done and *where* the data is stored is implicitly taken care of.

7.1 Query Driven Visualization

Automation is required by approaching the scalability problems of astronomical datasets from a visualization perspective. A high level of abstraction is necessary for interactive exploration through visualization, because this allows the visualization program to focus on what it does best: presenting data to the user. In particular, scalability should not be the responsibility of the software creating the visualization, but of the software that provides the necessary data.

With standalone visualization software, the user is required to deliver data manually. A more scalable method is to have the user request the data from within the visualization. Traditionally this requires the user to know where the data is stored and it is usually not possible to create new data this way.

In this thesis we introduce *query driven visualization*, a technique where data is requested directly from within the visualization in a way that allows users to not only request data in a declarative way, but also create new data. As a result, there is no difference between requesting existing data and creating new data.

Existing data can be discovered without having to describe where and how the required data is stored. New data is automatically created in an optimal way in case the requested data does not yet exist. These two ingredients provide scalability: processing is only spent on data that is required for the visualization and data is reused as much as possible. Neither the scientist nor the visualization software has to worry about how the data is processed or stored, because this is implicit.

A request for data can be formulated in an implementation independent way. This allows software to interoperate more easily than before, while at the same time the applications need less detailed knowledge about each other's inner workings. In the work done for this thesis, we used the *Simple Application Messaging Protocol*, a standard from the *International Virtual Observatory Alliance* to achieve this abstraction.

The research in this thesis primarily concerns astronomical catalogs. These catalogs can range in size from a full survey with billions of objects and hundreds of parameters, to a handful of objects with a few parameters used for a visualization. We also use the term *sources* to denote these objects, since most of them are detected due to the light they emit.

7.1.1 Automatic Data Discovery and Creation

Query driven visualization relies on the idea that all data handling is automatically performed in an optimal way. In this thesis we develop new methods to achieve this for astronomical catalogs.

An important aspect is that we make a distinction between a catalog itself and its contents. A catalog is created by defining what it contains in a conceptual way. The determination of the exact set of sources in the catalog, and the derivation of the actual values of their parameters is not necessarily performed at the time the catalog itself is created. Furthermore, the contents of the catalog are not necessarily stored once they are known.

This separation has enabled us to create mechanisms to handle data requests from

query driven visualization with the desired automatic scalability. These mechanisms can be summarized into four categories that roughly represent the chronological order in which they are used to deliver the required data for the visualization:

- *Catalog Discovery*: Catalogs that can be used to deliver or create the requested data are found automatically. Not only does this relieve scientists of managing the administration of their own datasets, but, more importantly, allows them to use data from collaborators automatically. Sharing data is therefore implicit.

  Discovery of catalogs is done by searching through the information that is stored about the catalogs; it is not necessary to inspect the catalog data itself. This is a necessity because the catalog can simply be too large and more importantly, because the actual contents of the catalog do not have to be stored at all.

- *Catalog Creation*: New catalogs are defined automatically if there are no existing catalogs that meet the requirements of the data request. Other datasets that are required for the creation of the new catalogs are recursively found or created in the same way.

  These new catalogs are created such that they are most suitable for reuse in possible future requests. This prevents the creation of many related and possibly overlapping catalogs, minimizing the required administration and duplicate datasets.

  For example, if a specific parameter has to be calculated for a small set of requested sources, then the catalog containing these parameters will be defined for the largest set of sources the calculation is applicable to. This way, the newly created catalog can be reused when the same parameter is later requested for a related, but slightly different, set of sources.

- *Catalog Processing*: The discovered or created catalogs can be processed partially. Only those parts of the catalogs that are required for the requested end result, are processed. Therefore, the processing is kept to a minimum, ensuring scalability.

  This is achieved by temporarily reorganizing the order of the operations that are required to produce the end result. For example, selections of sources and parameters are performed before the calculation of new parameters.

  The processing itself can be performed on different machines. Some operations, such as selecting subsets of sources, can be performed best on a database, while others, such as the calculation of parameters, can better be performed on a distributed computing cluster. For interactive visualization of small datasets it can be useful to process everything on the workstation of the scientist.

- *Catalog Storage*: The processed parts of the catalogs are stored only if this is necessary for performance. That is, they are stored in case the cost of reprocessing the data on demand is higher than the costs of the required storage.

  This is especially important for astronomical catalogs, because many operations to create catalogs, result in partial copies of existing catalogs, such as selecting

a subset of sources or parameters from another catalog. In the particular case of selecting a subset of sources, it is also possible to only store the identifiers of the selected sources, without creating copies of all their parameters.

The catalog data itself can be stored in a central database and/or on the local machine of the scientist. The first is essential to work with large datasets and for collaboration; the latter is essential for real time interaction.

A key element in all of this is that, from a user perspective, it is not necessary to know exactly how the data is stored, as long as it can be accessed in a feasible way. It is the data that matters, not how it is stored and scientists should not have to be encumbered with the task of managing the storage of data.

The same line of reasoning applies to processing. What is important is that the data is created according to the desires of the scientist, not exactly how the processing itself is performed.

7.1.2 Data Lineage

All the above mechanisms to handle astronomical catalogs in the most optimal way, depend on keeping track how catalogs are processed. Standards for storage of astronomical data have been invented long ago. The current development that plays a central role in this thesis is that not only the data itself but also information about its creation is stored digitally, called *data lineage*.

In this thesis we investigate how data lineage of astronomical catalogs should be treated, in particular with scalability and visualization in mind. There are two pieces of information that define a dataset: other datasets from which the new dataset is derived, and information about the used methods and chosen values for free parameters in these methods. In specifying this information it is not directly relevant *how* these methods are performed to create the described dataset.

The information about the creation of datasets is called *data lineage*, and defining the data lineage of a catalog is enough to create it. The data lineage of a catalog contains all the information required to process it, that is, it unambiguously defines how to create the contents of the catalog. Therefore it is not actually necessary to process the catalog.

Most of the problems that were encountered in designing query driven visualization mechanisms for astronomical catalogs, are solved by the use of data lineage. This is possible because the data lineage is more than just the creation history of data: it also facilitate parsing of data requests, discovery of data, creation and processing of data, etc.

Properly designing a structure for data lineage to achieve all of this is difficult. Maximizing the amount of information that can be inferred from the data lineage alone, without inspecting the described data itself, is one of the most important elements in this. The primary way this is accomplished in this thesis is by defining processing steps as elementary as is feasible. Complex processing steps can subsequently be defined as a composition of the more elementary ones. This results in

many intermediate datasets. However, this does not increase the required data storage, because with full data lineage it is not necessary to store the data of most of these intermediate products and because the same stored bit can be used in different datasets.

The more elementary the processing steps are, the better the above mechanisms can fulfill their task. In particular the reorganization of processing steps to process a catalog partially benefits from the elementary operations, because this allows the reordering to be more fine-grained.

7.1.3 Flexibility

Flexibility in processing is just as important as scalability in exploring data through interactive visualization. Attaining flexibility seems to be at odds with scalability and automation, because the latter often come with a loss of control. However, through query driven visualization the opposite is true: scientists have more control over how their data is derived, exactly because as much as possible is automated.

The automation requires a well defined structure on how to process data, through data lineage. This can be utilized in visualization. Firstly, the visualization software can use the data lineage within the visualization, by representing data in different ways based on how it is created. Secondly, the data creation process can be visualized itself. This makes it possible to inspect the creation of data from within the visualization, for example to determine whether the data is suitable for the specific goal that the scientist has in mind. Lastly, it is even possible to influence the creation of data from within the visualization in an abstract way.

7.1.4 Astro-WISE

Astro-WISE is an is an information system that is designed with the properties required for query driven visualization. Astro-WISE is created by the Astro-WISE Consortium, coordinated by OmegaCEN-NOVA. OmegaCEN is an astronomical data center for storage and processing of OmegaCAM optical wide field imaging surveys (such as KIDS). In recent years functionality has been added which made it a universal tool for astronomical data processing, archiving and scientific analysis. It is designed to accommodate raw data for thousands of nights of observations with more than a petabyte of data storage. It utilizes federated databases and data servers, and parallel compute clusters to manage these vast amounts of data. In this thesis we describe how query driven visualization of catalogs is implemented within Astro-WISE.

7.2 Galaxies and their Environment

Large catalogs are necessary to study the evolution of galaxies, and the research in this thesis offers a way to handle these. The universe contains billions of galaxies, that each consist of billions of stars: almost all sources that are visible with the naked eye belong to our own galaxy.

Generally speaking there are two kinds of galaxies: red elliptical galaxies and blue spiral galaxies. This is a too simplistic model of reality, even though many other properties of galaxies can be encompassed in this bimodality. For example, red spiral galaxies and blue elliptical galaxies exist as well.

Galaxies evolve during their lifetime and that the interaction with their environment plays an important role in this. Regions with a high density of galaxies, such as clusters, have a higher fraction of elliptical galaxies than regions with a low density. This is the so called morphology-density relation. A simplified model of the evolution of galaxies in relation with their environment, is that galaxies form as blue spiral galaxies in low density regions, get pulled into clusters due to gravity, and transform into red elliptical galaxies due to interaction with their environment.

The regions on the edges of clusters is of crucial importance in studying these galaxies, because most of their evolution happens in these regions. At the same time are these edges hard to describe, because they have complex shapes and the environment varies quickly as function of location. No perfect method to quantify the environment of galaxies is known. Different methods are compared in this thesis and a (for astronomers) new method is introduced. The methods each have their own strengths and weaknesses. Therefore it is essential that astronomers have access to different methods and chose the most applicable method based on their goals.

The set of galaxies that is used to quantify the environment is just as important as the choice of method. In this thesis we introduce a method to detect the edges of clusters by treating the presence of red and blue galaxies separately: the edges of clusters can be defined as regions with a high density of galaxies with a large contribution of blue galaxies. These edges of clusters contain more red spiral galaxies and more blue elliptical galaxies than is expected from the morphology-density relation alone. Two different morphological classifications are used to investigate this effect and in both cases there are more red spiral galaxies in these regions. However, these are also the regions where the differences in classification between the methods is the largest. This indicates that there are a lot of galaxies in this region of which the morphology is difficult to quantify. This could be because these galaxies are in the middle of an evolutionary transition.

7.3 Outlook

The query driven visualization and associated mechanisms as described in this thesis are only the beginning of what is possible. It is insightful to look how query driven visualization might develop in the future.

7.3.1 Data Lineage Standards

The benefits of the query driven way of interacting with data can be traced back to the idea that the data lineage provides a separation between what scientists want to do with their data and how this goal is accomplished. In designing data lineage, the focus should be on *what* the data represents and *what* is done with the data,

not *how* or *where* the data processing or data storage exactly takes place. On the long term this could lead to standardized descriptions of data processing, similar to standardized data formats, that can easily be shared between environments.

7.3.2 Ultra-fine Partitioning of Data Lineage

In this thesis we argue that processing steps should be as elementary as possible. This allows partial processing of data by reorganizing the required operations to create a dataset. This reorganization is done by local permutations, because this only requires knowledge of the commutation rules between processing steps. The more elementary the processing steps are, the easier it is to define the commutation relations between them and the more fine-grained this reorganization can be done. Ultimately it should be possible to partition processing steps to steps as small as individual arithmetic operations such as multiplications and additions.

Not only the data lineage should be defined in small parts, but also the data itself. In a sense, every pixel value or catalog element can be seen as a separate data product. Even properties like the number of sources in a catalog is a processing result in its own right. Such an extreme separation allows every component of a data product to be processed and stored on its own.

Nonetheless, it is not necessary to actually treat all these small pieces of information separately all the time. Large datasets can be split up when this is required automatically and the data itself can often be stored as part of the original whole. Vice versa, datasets can be combined into larger datasets without requiring extra data storage. As an extreme example this would make it possible to treat the images of survey as a large full sky image, while retaining the possibility of handling individual pixels at the same time.

7.3.3 Incremental Visualization

In this thesis we have restricted ourselves to requesting a complete dataset for visualization. That is, the scientist defines which visualization is required and the required data is delivered automatically, but always as a whole. Creating a catalog it its entirety can still be very time-consuming, even if it is created in the most efficient way.

This problem can be resolved by *incremental visualizations*. That is, the visualization software will receive the requested data in parts and build the visualization in steps. Query driven visualization is an excellent mechanism for this, because data is already processed in parts. There are two main ways this can be achieved with catalog data. The simplest way is to derive the required catalog data for an growing set of sources until the catalog has been created for the entire requested set of sources. The second way is to derive the parameters of the sources in varying precision. That is, the parameters are first derived with a few digits precision and more digits are calculated as time goes on, refining the visualization.

This will improve the response time for interactive visualization, making interactive exploration of data faster as a result. Scientist can get an overview of their

requested visualization quickly and can either decide to wait for the details to improve, or decide to cancel the operation. Furthermore, it would be possible to vary the details of the visualization. For example by spending more time on the processing of outliers in the visualization.

There are several challenges that have to be overcome to achieve this. Most importantly, the visualization software should be able to handle data that changes in precision. Furthermore, the user should be able to specify interactively which parts of the visualization require more detail. This requires on the fly reevaluation of how the data should be processed. The data processing should not only be able to work on the level of individual parameters of sources and individual pixels as described above, but even on individual bits of these values. Most of the solutions to these challenges can be seen as a continuation of the mechanisms underlying query driven visualization, and have therefore become within reach by the work described in this thesis.

7.3.4 Current and Future Astronomical Projects

Several large astronomical projects are currently under way or planned for the future. The OmegaCAM optical camera will soon start operations on the VLT Survey Telescope, resulting in KIDS, a 1500 square degree survey. This will be the optical counterpart of the infrared survey VIKING from the VISTA telescope. The future Euclid mission plans to cover 20 000 square degree and will detect up to 10^{10} galaxies. In this thesis we primarily discussed the use of query driven visualization with released survey data. There are several ways in which these ongoing projects can benefit form query driven visualization, in particular during their earlier phases.

The commissioning, operations and data reduction of these new instruments is done in collaborations between scientists in different institutes. In these early phases there is a need for close interaction with the data to experiment with process parameter settings and to explore different data reduction methods. Query driven visualization makes this experimentation faster. Furthermore it becomes easier to gain insight in the developments of other members of the collaboration and to integrate their results in ones own work.

The focus in this thesis is on query driven visualization of catalog data. Survey operations would benefit from combining this with image handling. For example this would make it trivial to answer questions such as “how would this color-color plot change if a different overscan correction method is used during the image reduction?”.

Other astronomical projects with large datasets can also make use of query driven visualization. The LOw Frequency ARray (LOFAR) project produces such enormous amounts of raw data that most of it cannot be stored. LOFAR uses a huge amount of low-cost sensors that are placed in stations spread over an area of 1500 km wide. Data processing and storage is performed both locally at the stations, at a central location and on distributed grids. It is a nontrivial question to decide what processing should take place where and which parts of the data should be stored and for how long.

These are the exact same problems that arise with query driven visualization. Therefore it would be interesting to see whether the same solutions can be applied.

That is, whether the LOFAR processing and storage challenges can be tackled in a dynamical way by starting at the desired end product.

7.3.5 Galaxy Evolution

The large datasets required for studying the properties and evolution of galaxies formed the rationale for our development of query driven visualization and its underlying mechanisms. Only a fraction of the field is covered by the research in this thesis, and there are many more questions. Query driven visualization can be used to answer them with data from current surveys such as KIDS, performed with the OmegaCAM on the VST and far future surveys planned with the Euclid satellite.

The relations between the properties of galaxies are intertwined. Many galaxy properties correlate to some extent with each other. Query driven visualization can help disentangle these different relations by combining it with software for automatic detection of trends, clusterings and outliers in the data. This software can use the same query driven mechanisms to achieve scalability. Coupling automatic analysis with visualization will combine computing power with human intelligence, allowing for efficient exploration of this multi-parameter space. For example, there is a large parameter space to explore in color space alone. In this thesis we focused on five optical bands, resulting in 10 different colors. Combining this optical data with another five infra red bands results in 45 different colors. It will be interesting to explore data from different wavelength regimes.

Query driven visualization is useful in comparing different methods to quantify properties of galaxies and studying their effects, because it allows flexible exploration of methods and parameter settings. Adding more colors to the parameter space opens up the 'method' space as well, because many of the properties of galaxies, such as mass, can be derived from their colors. Applying our new techniques we discovered that the color of galaxies can be used in quantifying environment. There is no a priori reason to assume that the specific color that was used $(u - r)$ is the most suitable color to use. Therefore, the use of other colors and other galaxy properties in the quantification of environment should be investigated.

In this thesis we present a mechanism to detect the edges of clusters, a particularly difficult to quantify environment. We demonstrated that these regions are interesting for galaxy evolution. For example, they contain larger fractions of red spiral galaxies and blue elliptical galaxies than other regions. The next step is to investigate whether it is possible to associate evolutionary stages and mechanisms to these regions and galaxies within them. For this it is required to study these regions in more detail, and with more information such as colors from different wavelength regimes or spectral data.

Furthermore, these edge regions contain many galaxies for which the morphology apparently is hard to quantify, because they are classified differently by different methods. A reason for this could be that these methods actually quantify different properties of galaxies, because describing morphology with a single parameter might not be feasible at all. This leads to the interesting question of what would be a proper way to study the morphology of galaxies and query driven visualization can help to

explore the possibilities.

7.4 Conclusions

Query driven visualization allows scientists to discover existing datasets and create new datasets by requesting data directly from within the visualization. New datasets are automatically created in such a way that they are most suitable for reuse in future requests, preventing duplications of data. The subsequent processing of the datasets is limited to those parts that are necessary to create the data for the requested visualization, achieving implicit scalability.

The same mechanisms ensure that scientists have control over the methods and parameters that are used to process their data, achieving flexibility. This allows scientists to interact with their data in a conceptual way and allows them to focus on *what* they want to do with the data, because *how* the processing is performed and *where* the data is stored is implicitly taken care of.

This is achieved by storing how data is processed. The processing should be defined in steps that are as small as feasible. This maximizes the amount of knowledge that can be inferred from the processing information without having to inspect the data itself. This in turn, makes it possible to automate administrative tasks such as the discovery and scalable creation of data.

Applying our techniques we discover that the edges of galaxy clusters can be detected by taking color into account in the quantification of environment. We detect an overabundance of red spirals and blue galaxies in these regions, indicating evolution of galaxies. Furthermore, these regions have a large fraction of galaxies for which there is a discrepancy in the classification of morphology through different methods, suggesting that these galaxies are in the process of morphological transition.

The current wide field surveys such as KIDS, VISTA and those planned for Euclid, will make it possible to study such evolution of galaxies in great detail, due to the billions of galaxy that they will detect. Other astronomical projects such as LOFAR will process even more data, part of which has to be done in real time. All these projects require the scalability and flexibility provided by query driven visualization and the underlying mechanisms. Ultimately, query driven data visualization is not only a bright possible future, but perhaps even an inevitable one.